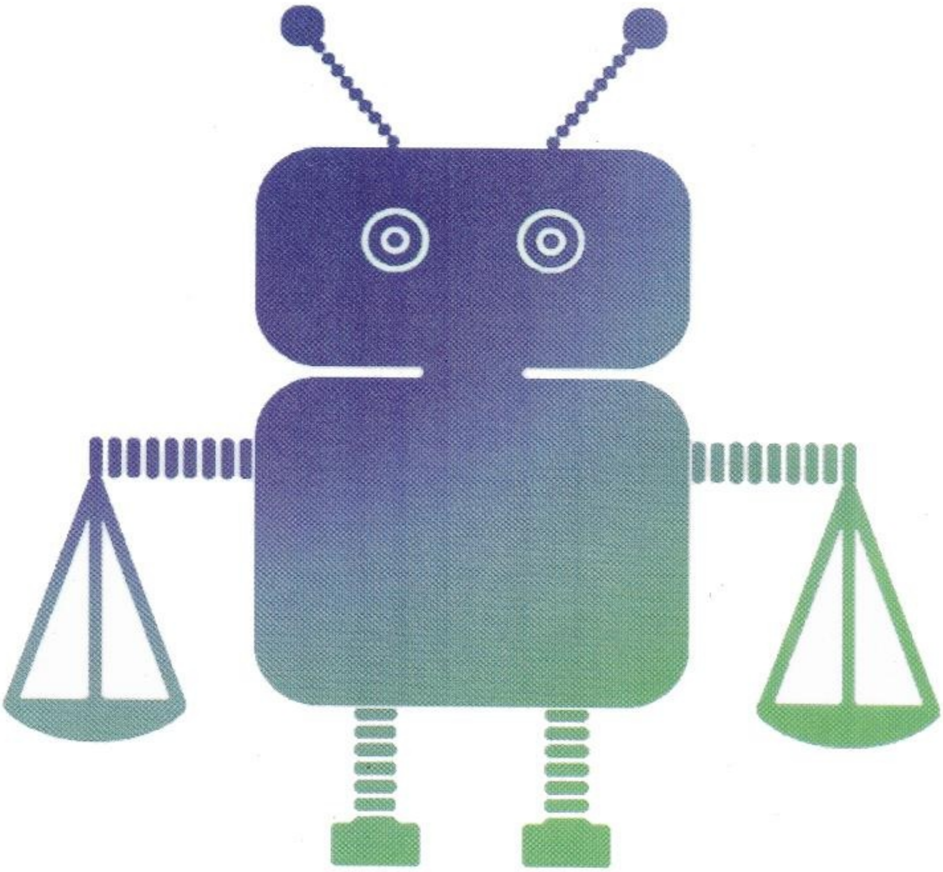
LA ÉTICA DE LOS ROBOTS

Mark Coeckelbergh



¿Tienen agencia moral los robots? ¿Pueden ser considerados responsables de sus acciones? ¿Nos reemplazarán en el trabajo? ¿Les debemos algo los seres humanos? Estas son algunas de las cuestiones de carácter ético y moral a las que tenemos que enfrentarnos en un contexto en el que el uso de robots y otros dispositivos inteligentes cada vez más sofisticados se ha vuelto habitual. En este libro, planteado como una guía ante este escenario inédito, Mark Coeckelbergh aborda directamente estas preguntas, para lo cual tiene en cuenta una gran variedad de dispositivos —desde el robot de compañía hasta el dron de combate—, sosteniendo en todo momento una premisa esencial: los problemas de la robótica no atañen solo a las máquinas, sino a la esencia misma de los seres humanos.

Mark Coeckelbergh

La ética de los robots

ePub r1.0

Titivillus 01.03.2025

Título original: *Robot Ethics*
Mark Coeckelbergh, 2022
Traducción: Lucas Álvarez Cangas

Editor digital: Titivillus
ePub base r3.0 (ePub 3)

Por la naturaleza de la red, puede que algunos de los vínculos a páginas web en el libro ya no sean accesibles.

Índice de contenido

Índice de contenido

CAPÍTULO PRIMERO

Introducción: ¿de qué trata la ética de los robots?

En la película de ciencia ficción estadounidense de 2004 *Yo, robot*, los robots humanoides sirven a la humanidad. Sin embargo, no todo va bien. Tras un accidente, un hombre es rescatado del coche que se está hundiendo por un robot, pero una niña de doce años no se salva. El robot calculó que el hombre tenía una probabilidad de supervivencia más alta: los humanos puede que hubieran hecho otra elección. En la película, los robots intentan más tarde tomar el control del poder que tenían los humanos. Los robots están controlados por una inteligencia artificial (IA), VIKI, que decide que restringir el comportamiento de los humanos y matar a algunos de ellos asegurará la supervivencia de la humanidad. La película ilustra el miedo a que los robots humanoides y la IA tomen el control del mundo. También apunta a hipotéticos dilemas éticos diferentes en el caso de que los robots y la IA llegasen a alcanzar inteligencia en un sentido general. Pero ¿es de esto de lo que trata o debería tratar la ética de los robots?

¿ESTÁN LOS ROBOTS DE CAMINO O YA ESTÁN AQUÍ?

Normalmente, cuando la gente piensa sobre los robots, la primera imagen que nos viene a la cabeza es la de un robot altamente inteligente con forma humana. A menudo esta imagen proviene de la ciencia ficción, donde encontramos robots que se parecen y se comportan más o menos como los humanos. Muchas narraciones nos alertan de que los robots tomarán el poder. El miedo que estas nos producen consiste en que dejarán de ser nuestros sirvientes y pasarán a convertirnos en *sus* esclavos. El mismo término «robot» significa «trabajo forzoso» en checo y aparece en la obra de teatro *R. U. R.* de Karel Capek, puesta en escena en Praga en 1921: justamente hace cien años. La obra es parte de una larga historia de narraciones sobre máquinas rebeldes similares a los humanos, desde el *Frankenstein* de Mary Shelley a películas

como *2001: Una odisea del espacio*, *Terminator*, *Blade Runner* y *Yo, robot*. En el imaginario público, los robots son, a menudo, objeto de miedo y fascinación al mismo tiempo. Tenemos miedo de que tomen el poder, pero, al mismo tiempo, resulta emocionante pensar sobre la creación de un ser artificial que sea parecido a nosotros. Como parte de nuestra herencia romántica, los robots son proyecciones de nuestros sueños y pesadillas sobre crear un otro artificial[1].

En primer lugar, estos robots son ante todo terroríficos: son monstruos y causan asombro. Pero al comienzo del siglo XXI surge en Occidente una imagen diferente de los robots: el robot como compañero, amigo y, quizás, incluso pareja. La idea es ahora que los robots no deberían estar limitados a las fábricas industriales o a lugares remotos en el espacio. En el imaginario contemporáneo, están liberados de su sucio trabajo esclavo, y entran al hogar como compañeros sociables útiles, agradables y, a veces, sexis, con los que se puede hablar. En algunas películas, al final se rebelan (piénsese en *Ex Machina*, por ejemplo), pero, generalmente, se convierten en lo que los diseñadores de robots llaman «robots sociales». Están diseñados para mantener una interacción «natural» humano-robot: esto es, para interactuar de la forma en la que estamos acostumbrados con otros seres humanos o mascotas. Están diseñados para no dar miedo o parecer monstruosos, sino para ser bonitos, útiles, entretenidos, divertidos y seductores.

Esto nos conduce a la vida real. Los robots no están de camino: ya están aquí. Pero no son exactamente como los robots que nos encontramos en la ciencia ficción. No son como el monstruo de Frankenstein o Terminator. Son robots industriales y, a veces, «robots sociales». Sin embargo, estos últimos no son tan inteligentes como los humanos o sus parientes de la ciencia ficción y, a menudo, no tienen forma humana. Incluso los robots sexuales no son tan inteligentes o tan conversacionalmente capaces como el robot representado en *Ex Machina*. A pesar de los avances recientes en IA, la mayoría de ellos no se parecen a los humanos en ningún aspecto. Dicho esto, los robots están aquí, y están aquí para quedarse. Son más inteligentes y tienen mayor capacidad de funcionar de forma autónoma que antes. Y tienen aplicaciones más orientadas al mundo real. Los robots no se usan solamente en la industria, sino también en la atención sanitaria, el transporte y la ayuda doméstica.

> Los robots no están de camino: ya están aquí. Pero no son exactamente como los robots que nos encontramos en la ciencia ficción.

A menudo, esto hace que la vida humana sea más fácil. Sin embargo, también hay problemas. Algunos robots pueden ser ciertamente peligrosos, no porque puedan intentar matar o seducirte (a pesar de que los «drones asesinos» y los robots sexuales también están en el menú de la ética de los robots), sino, normalmente, por razones mucho más mundanas, como que puedan quitarte el trabajo, puedan engañarte y hacer que pienses que son una persona, y puedan causar accidentes cuando los usas de taxi. Miedos tales como estos no son ciencia ficción: pertenecen al futuro próximo. Dicho de manera más general: desde el impacto de las tecnologías nuclear, digital y de otro tipo en nuestras existencias y en el planeta, hay una conciencia y un reconocimiento cada vez más presente de que las tecnologías conllevan cambios fundamentales en nuestras vidas, sociedades y entorno, y que, por lo tanto, mejor es que pensemos más, y de forma más crítica, sobre su uso y su desarrollo. Hay aquí una cierta urgencia: debemos comprender y evaluar las tecnologías ahora, antes de que sea demasiado tarde, esto es, antes de que conlleven impactos que nadie quiera. Este argumento también se puede utilizar con respecto al desarrollo y uso de la robótica: consideremos las cuestiones éticas planteadas por los robots y su uso en la fase de desarrollo en lugar de después de los hechos. Diré ahora más sobre los objetivos y el alcance de este libro.

OBJETIVOS DE ESTE LIBRO: CUESTIONES ÉTICAS Y REFLEXIÓN FILOSÓFICA

En la medida en que las tecnologías robóticas y de automatización abandonen el reino de la ciencia ficción y penetren cada vez más en nuestras vidas cotidianas, es importante no solo ver los beneficios y oportunidades potenciales, sino también discutir las cuestiones éticas y sociales que

plantean, ahora y en el futuro cercano. Considérese, por ejemplo, los robots industriales que son cada vez más inteligentes y trabajan con humanos en fábricas, los robots usados por usuarios vulnerables como es el caso de los niños, los coches autónomos que están siendo desarrollados por prácticamente todos los fabricantes de coches, los robots usados para cirugía en hospitales y los drones letales que se usan como armas de guerra. Este libro responde a estas cuestiones ofreciendo una concepción general de algunos problemas éticos y sociales clave junto con una gama de herramientas conceptuales de la ética de los robots para pensar sobre estos problemas. Muestra ejemplos de aplicaciones robóticas prácticas, identifica problemas éticos y sociales que plantean estas aplicaciones, y ofrece herramientas conceptuales para lidiar con estos problemas. La idea que está detrás de este ángulo práctico es que, de esta forma, la ética de los robots puede ayudar a hacer que el desarrollo y gobierno de la robótica procedan de una forma más éticamente responsable.

Además, el libro busca ofrecer una reflexión filosófica sobre qué son los robots y qué significa pensar sobre robótica para pensar sobre lo humano. Por ejemplo, nos hace reflexionar sobre cómo pensamos los seres humanos acerca del estatus moral. Esto proporciona una perspectiva más amplia que es muy necesaria en debates que normalmente están limitados a una preocupación ética más inmediata, y ayuda a explicar la profunda fascinación que ejercen sobre el público general los robots y las máquinas. Enseguida diré algo más sobre el significado filosófico de la ética de los robots. Pero haré primero una pregunta sencilla: ¿qué se quiere decir con ética de los robots?

¿QUÉ SON LOS ROBOTS? ¿QUÉ ES LA ÉTICA? Y «QUÉ QUIERE DECIR «ÉTICA DE LOS ROBOTS»?

Si queremos discutir en este libro sobre la ética de los robots, ¿de qué queremos hablar? ¿Qué es «robot», y qué es «ética»? Esto ya nos conduce a problemas de definición, una preocupación típica de filósofos.

En primer lugar, ¿qué se quiere decir con el término «robot»? Ya he mencionado su significado etimológico, relacionado con sirvientes y esclavos. Sin embargo, cuando tratamos de habérnoslas con los robots en la práctica, este es solamente un significado entre muchos. El significado del término es controvertido. Ni los expertos en robótica ni los filósofos se ponen de acuerdo sobre la definición de robot. Por ejemplo, el Instituto Internacional de Ingenieros Eléctricos y Electrónicos (Institute of Electrical and Electronics Engineers) define los robots como sigue: «un robot es una máquina autónoma

capaz de detectar su entorno, llevando a cabo computaciones para tomar decisiones y realizando acciones en el mundo real[2]». Pero ¿hace esto que un termostato sea un robot? ¿Y un lavavajillas? ¿Y un piloto automático? ¿Deberían los robots ser capaces de moverse? ¿Dónde está la línea que los separa de otros tipos de *hardware*? ¿Es un coche autónomo un robot? ¿Cómo de material tiene que ser? Los robots tienen componentes de *hardware* (físicos) y de *software* (código). Si un agente artificial consistiera solo en *software*, se llamaría un «bot»: no se consideraría un robot. Patrick Lin y sus colegas han argumentado que, en contraste con los ordenadores o la IA no corporeizada, un robot «puede ejercer influencia sobre el mundo de forma directa»[3]. Aun así, ¿por qué es menos directa la influencia de un «bot»? ¿Y qué sucede si el *software* está conectado al *hardware*, sin tener la forma humana o animal que normalmente asociamos con «robot»? Finalmente, ¿cómo de autónomo e inteligente debería ser un «robot»? Por ejemplo, a veces los robots se combinan con IA, pero esto no tiene por qué ser necesariamente así. Dada la importancia cada vez mayor de la interactuación entre los robots y los humanos, un término usado habitualmente es el de «interacción humano-robot»: el énfasis no está en el robot como un artefacto material, sino en la interacción entre humanos y robots. Hay definiciones amplias y estrechas de robots. En este libro, me centraré en aquellos robots que tienen *hardware*, y resaltaré a menudo problemas éticos planteados por robots que tienen un alto nivel de inteligencia, autonomía e interactividad, pero también incluiré otros tipos de robots.

Sin embargo, las definiciones técnicas no son suficientes. La controversia sobre el futuro de la robótica y las imágenes de la ciencia ficción apuntan tanto al contexto del robot, así como a las dimensiones humanas, sociales y culturales de la robótica. Lo que un robot «es» no puede ni debe ser reducido al artefacto material «robot», sino que debe conectarse más bien con su uso, y sus contextos sociales y culturales. Lo que un robot «es» siempre está moldeado por el uso humano, (inter) acción, subjetividad y cultura. Por ejemplo, si ciertos humanos particulares usan un robot particular de una forma en la que lo tratan como una mascota, el significado de ese robot en ese contexto y situación no está descrito exhaustivamente llamándolo cosa o máquina. Las narrativas sobre robots mencionadas al comienzo de este capítulo no son irrelevantes a la hora de pensar sobre los robots: es mejor que les prestemos atención, ya que influyen no solo en la percepción pública de la robótica, sino también en su desarrollo. Los ingenieros y diseñadores no son

inmunes a la ciencia ficción: también les influye *a ellos*. Algunos sueñan con construir una máquina con forma humana o, como usuarios del robot, proyectan todo tipo de significados antropomórficos (con forma humana) sobre el robot. Algunos evocan tales significados a propósito para hacer que su robot sea más «social» y aceptable. Usando de nuevo la imagen anterior, dar a los robots un nombre personal (el nombre de un humano o animal) no es una excepción: es una práctica común. Las definiciones técnicas de los robots, por lo tanto, son necesarias, pero no suficientes en la ética de los robots. Los robots no son simples máquinas: siempre son al mismo tiempo humanos, sociales y culturales. Su significado no puede reducirse a definiciones técnicas.

> Lo que un robot «es» no puede ni debe ser reducido al artefacto material «robot», sino que debe conectarse más bien con su uso y sus contextos sociales y culturales.

Los robots comparten este sentido multifacético con otras tecnologías. Así como la tecnología en general, el término «robot» se puede referir a una variedad de fenómenos. La tecnología no trata solo sobre artefactos materiales. También trata sobre el conocimiento necesario para desarrollar y usar estos artefactos, así como sobre la ciencia relacionada; normalmente de los programas de computación inmaterial; actividades como el uso, diseño y mantenimiento, y de los actores humanos y sociales involucrados en el desarrollo y uso de las tecnologías. Su despliegue tiene lugar dentro de un marco socioeconómico concreto (por ejemplo, el capitalismo) y en una cultura particular, su uso y desarrollo implican suposiciones sobre lo humano y la tecnología (por ejemplo, que los robots son simples herramientas, que son lógicos y racionales, etc.), y la tecnología trata (como ha mostrado Martin

Heidegger en su famoso ensayo sobre tecnología) acerca de nuestra actitud hacia nuestro mundo y la concepción que tenemos de él[4]. Por lo tanto, los «robots» evocan todos estos significados y contextos de la tecnología. Son artefactos, pero su desarrollo y uso también requiere de tipos particulares de conocimiento y ciertas capacidades. Además, existe la ciencia de la robótica. E l *hardware* del robot es material, pero los robots también tienen componentes inmateriales (código). Están vinculados con muchas actividades que realizan los seres humanos, como el diseño de los robots, así como el uso y la interacción con ellos. Su desarrollo involucra tanto a actores sociales como a las empresas y el Estado. Se desarrollan y se usan en sociedades capitalistas como los Estados Unidos, pero también en otras distintas. Pueden ser parte de la cultura occidental, o su desarrollo puede tener otras raíces culturales. Su desarrollo acepta normalmente un modelo de lo humano (por ejemplo, el humano como máquina) o de humanos particulares (por ejemplo, ancianos infantilizados). A menudo están conectados con una concepción particular del mundo, por ejemplo, la concepción según la cual la naturaleza está ahí para ser explotada, o que los consumidores son una especie de ganado que proporciona información para la economía de datos. Como muestran estos ejemplos, definir a los robots implica ya concepciones particulares, incluyendo concepciones *normativas*, y que estas ya están *haciendo* ética de los robots y «filosofía robótica»: todo gira alrededor de comprender y evaluar a los robots junto con (su relación con) los humanos. De forma más general, el lenguaje que usamos para hablar sobre robots (como desarrolladores, usuarios o, ciertamente, como filósofos) tampoco es neutral. Lo que un robot «es» depende de cómo hablemos *con y sobre* el robot[5].

En segundo lugar, ¿qué es la «ética»? La ética es en sí misma un término controvertido que también puede tener muchos significados. Los filósofos tienden a estar de acuerdo en que tiene que ver con preguntas normativas, particularmente con qué *debemos* hacer y cómo *debemos* vivir. La ética se puede referir a los principios y valores morales, pero también a la rama de la filosofía académica denominada «ética», que discute los principios, teorías y conceptos éticos. Algunos filósofos (por ejemplo, los pragmatistas) ponen más énfasis en la práctica ética y en la experiencia moral que en principios éticos. Dependiendo de la teoría moral usada, la ética se puede centrar en las obligaciones morales, las consecuencias morales, el carácter moral u otros elementos. Además, la «ética» puede ser entendida como una cuestión sobre límites (para evitar el comportamiento inmoral), pero de forma más positiva,

puede formularse como una cuestión sobre la vida buena. Puede tratar sobre lo que es bueno para los individuos, así como lo que es bueno para la sociedad. El término «ética» normalmente se refiere a la ética que concierne a los humanos, sin embargo, también se puede referir a éticas que conciernen a no humanos como animales o, ciertamente, robots. «Ética» puede significar ética dirigida *hacia* los humanos, animales, y así sucesivamente (son el objeto de la ética), o puede significar ética *de* o *para* los humanos y otras entidades (y son, en ese caso, el tema de la ética).

En tercer lugar, la expresión «ética de los robots» se puede referir a la ética de cómo los *humanos* deberíamos usar, interactuar y desarrollar robots de una forma que conduzca al bien para los humanos u otras entidades, tales como animales o, quizás, incluso robots (los humanos son los sujetos éticos; los robots son los medios para alcanzar el fin de la ética humana), o se puede referir a una ética para *robots*. Aquí se usa a menudo el término «ética de las máquinas»; nos referimos en este caso a la ética que puedan tener los robots: los robots son vistos como (potenciales) sujetos éticos. Es importante no confundir estos significados diferentes. La ética de los robots *no* trata necesariamente sobre «dotar a los robots de una ética» o de que «los robots tengan una ética». Por ejemplo, Peter Asaro distingue entre tres significados de ética de los robots: «los sistemas éticos incorporados en los robots, la ética de los que diseñan y usan los robots, y la ética de cómo debemos tratar a los robots[6]». El primer significado tiene que ver con la ética de los robots en tanto sujetos éticos, como cuando se dice que los coches autónomos deberían tener una ética incorporada. El segundo y tercer significados tienen que ver con la ética de los humanos en tanto sujetos éticos, como cuando se discute si los que diseñan robots deberían ser corresponsables de su uso (por ejemplo, para propósitos comerciales o en una guerra) o se afirma que los robots no deberían ser «torturados».

Hablando de forma descriptiva, la ética tiende, al menos parcialmente, a ser diferente entre culturas y sociedades e, incluso, dentro de las propias sociedades. La gente puede tener diferentes opiniones éticas sobre cuestiones particulares. No todo el mundo está de acuerdo con cómo gobernar o dirigir el comportamiento de la propia vida, y lo que es mejor para la sociedad. La ética en este sentido descriptivo ha variado históricamente. Por ejemplo, muchos de nosotros tendemos a pensar que los animales tienen algunos derechos y, de forma más general, un estatus moral más elevado que las cosas. Esta variación cultural e histórica es también relevante para la ética de los robots:

las actitudes éticas hacia los robots y las creencias sobre ellos tienden a diferir entre culturas, y pueden variar en el tiempo. Japón se pone a menudo como ejemplo de cómo varían las actitudes hacia los robots. Frecuentemente se dice que la gente en Japón tiene una actitud más positiva hacia los robots debido a su específica historia cultural y popular (con robots representados como ayudantes de los humanos) y la continua influencia de concepciones del mundo en las que los no humanos pueden tener un estatus espiritual[7]. Algunas personas sostienen que en el futuro otorgaremos o deberíamos otorgar derechos a los robots.

Nótese que en este libro usaré el término «ética» y «moralidad» de forma indistinta. Sin embargo, quienquiera que reflexione sobre ética de los robots es libre de distinguir entre estos términos si se muestra que puede ganarse algo con tal distinción. Nótese también que, en términos académicos, el campo de la ética de los robots puede ser visto como una rama de la ética (aplicada) y de la filosofía práctica. Pero se puede relacionar también con el campo de la filosofía de la tecnología. Es, por lo tanto, vista como una de las tecnologías a las que se les puede aplicar conceptos y teorías de la filosofía de la tecnología.

En este libro usaré «ética» y «ética de los robots» en muchos sentidos diferentes. En cada capítulo o argumento haré explícito el tipo de «ética» de la que se está hablando. Por ejemplo, el libro incluirá discusiones sobre si los robots pueden ser agentes morales (sujetos éticos) y pacientes morales (objetos éticos). No solo se tratará de la ética como preocupación por hacer lo correcto, sino también de la ética de la vida y de la sociedad buenas. Estas elecciones y mi tratamiento de estas cuestiones reflejan de forma inevitable mi propia concepción de cómo hacer ética. Por ejemplo, en contraste con muchos libros de texto sobre ética, lo predeterminado no es centrarse en la ética del individuo. El capítulo siguiente se hace preguntas concernientes al impacto sobre la sociedad en lugar de solamente sobre los individuos. Además, en contraste con muchos libros de ética aplicada, su proyecto no está centrado en la aplicación de una teoría moral normativa. El lector no encontrará capítulos con nombres como «deontología», «consecuencialismo» o «ética de la virtud». En su lugar, el libro toca diversas cuestiones éticas, como la responsabilidad, la dignidad humana y las cuestiones concernientes al estatus moral. Estas discusiones implicarán otros conceptos y teorías, incluyendo, por ejemplo, una teoría sobre la responsabilidad, así como una teoría moral normativa, como el consecuencialismo (capítulo 5) y la ética de

la virtud (capítulo 6). Pero el *punto de partida,* no es la teoría como tal: son las cuestiones éticas que se plantean desde el uso y la práctica tecnológicos.

No todos los expertos en ética de los robots estarán de acuerdo con este enfoque ni con las definiciones que se han proporcionado. Mi presentación de lo que queremos decir con robots, ética y ética de los robots puede (como todas las definiciones de conceptos filosóficos) ser discutida. Otros pueden poner mayor énfasis sobre la ética individual o las teorías morales. Por ejemplo, se puede encontrar otro tratamiento en la obra de Keith Abney, quien responde a la cuestión de lo que es la ética de los robots discutiendo no solo qué son la moralidad y la ética, sino también qué queremos decir cuando hablamos de derechos y obligaciones morales, cuáles son las principales teorías morales contemporáneas y cómo prestan apoyo la ética de los robots, y qué es una persona[8]. Diré algo más sobre el enfoque de mi libro.

ENFOQUE, ESTRUCTURA Y ALCANCE DEL LIBRO

Mientras que este libro se centra en introducir conceptos y teoría filosófica en el campo de la ética de los robots, también busca hacerlo de una forma que muestre su relevancia para las cuestiones del mundo real. Para esto, liga ciertas aplicaciones robóticas específicas con discusiones filosóficas. Por ejemplo, conecta los coches autónomos con las cuestiones que conciernen a la responsabilidad moral. Así, introduce al lector al primer y más conocido objetivo, así como a una definición de la ética de los robots: contribuir (mediante conceptos y deliberación filosóficos) a la comprensión y al tratamiento de *cuestiones éticas con robots en el mundo real*. Este es un objetivo importante que puede, y tiene que, continuar siendo adicionalmente promovido mediante investigación inter- y transdisciplinar, implicando en ello diálogos con, por ejemplo, los profesionales de campos como la ingeniería, la política y el derecho. Cada vez más gente, dentro y fuera de la academia, se ve atraída por este emocionante proyecto.

Este énfasis sobre las cuestiones éticas del mundo real con robots actuales y de un futuro próximo tiene como consecuencia el que se ponga menos atención a temas como la superinteligencia (esto es, máquinas con una inteligencia hipotética que excede por mucho la humana), o máquinas que tengan inteligencia artificial general (la inteligencia hipotética de una máquina que puede entender y aprender cualesquiera tareas intelectuales que pueda acometer un humano). En mi opinión, los debates sobre estos temas nos desvían a menudo del tratamiento de los problemas del mundo real y del

futuro cercano en el ámbito de la ética de los robots. Dicho esto, reconozco sin reservas que tales discusiones pueden contribuir a el (los) objetivo(s) filosófico(s) de la ética de los robots (ver más abajo). Por lo tanto, en el último capítulo, introduzco el tema de la superinteligencia en el contexto de los enfoques transhumanistas de la ética de los robots.

Además, mi énfasis sobre la ética de los robots en el mundo real, más bien que sobre los robots de ciencia ficción, no implica que crea que la ciencia ficción sea completamente irrelevante para la ética de los robots. Según ya he sugerido, es relevante estudiar cómo se imaginan, y han sido imaginados, los robots en la ciencia ficción, cómo esta imaginación y estas narrativas influyen sobre cómo los robots se perciben y se usan hoy en día, y cuáles son las implicaciones normativas. He contribuido a este tipo de trabajo con mi libro *New Romantic Cyborgs*. La ciencia ficción también ofrece experimentos mentales interesantes para los filósofos. Uso algunos de ellos en este libro, especialmente en mis introducciones a los distintos capítulos, cada uno de los cuales se refieren con frecuencia a películas de ciencia ficción. Y algunas veces pueden extraerse lecciones éticas y políticas de la ciencia ficción. Por ejemplo, como ha mostrado Eileen Hunt Botting, *Frankenstein* puede ser un recurso para pensar sobre la responsabilidad y los derechos[9]. En el contexto de nuestro tema, esto podría significar que la historia nos ayude a hacernos preguntas sobre la responsabilidad que conlleva la creación de robots. ¿Es aceptable que la gente que crea y vende robots «abandone» sus creaciones, y deje los desafíos éticos que plantean en manos de los usuarios y su sociedad, o deberían asumir su responsabilidad por tales creaciones? Para muchos expertos contemporáneos de la ética de los robots, esta es una pregunta retórica: creen que los diseñadores y desarrolladores de la tecnología, entre otros, deberían ser responsables de la tecnología (ver también el sumario que presento de la investigación y la innovación responsables en el capítulo 4). Por el contrario, para aquellos que están involucrados en robótica y que aún no se han tomado en serio la ética de los robots, o para aquellas personas procedentes de las humanidades que aún no han conectado con la tecnología contemporánea, reflexionar sobre la responsabilidad de los ingenieros bajo la luz de *Frankenstein* puede ofrecer un buen punto de partida para la reflexión y, así, contribuir al principal objetivo de la ética de los robots: tratar los problemas éticos de la robótica en el mundo real.

Sin embargo, en mi opinión, la ética de los robots debería tener un segundo (pero no secundario) objetivo, que recibe menos atención, pero que

quizás nos ofrezca una explicación más profunda de por qué el público más general está tan interesado en este campo, a saber: contribuir a la *filosofía*, y no solo a la reflexión sobre la tecnología, sino, por ejemplo, también a la reflexión sobre lo humano. Reflexionar sobre los robots no es solamente pensar sobre los robots, sino que también toca muchas otras cuestiones que tienen una relevancia filosófica más amplia. Es por esto por lo que a lo largo del libro la ética de los robots está conectada con cuestiones de otros subcampos de la filosofía. Por ejemplo, muestra que la investigación sobre la ética de los robots conduce a cuestiones ontológicas (ver la sección previa sobre lo que es un robot), epistemológicas (por ejemplo, qué es la pericia, y qué tipo de conocimiento tenemos sobre el estatus ontológico y moral de otras entidades) y filosófico-antropológicas (qué son los humanos y qué significa ser humano). El libro muestra que preguntarse sobre los robots es también siempre preguntarse sobre los humanos junto con su moralidad, sus prácticas y sus instituciones: esto se muestra en cada capítulo, y finaliza con una reflexión sobre las relaciones entre los robots y los humanos. Por supuesto, reflexionar sobre los robots nos ayuda a comprender y evaluar mejor la (una particular) tecnología. Como tal, es parte de la filosofía de la tecnología y de la ética de la tecnología. Pero también trata sobre los humanos. Un tema común que recorre este libro es que los robots funcionan como herramientas para comprendemos mejor a nosotros mismos: herramientas usadas por los científicos para comprobar y mejorar sus modelos sobre los seres humanos y otros seres naturales, pero, además, herramientas usadas por los *filósofos* para reflexionar sobre lo que es ser humano. De ahí que el título inicial de este libro fuera *Robotic Mirrors* (Espejos robóticos): usamos los robots como espejos para reflexionar, entre otros objetivos, sobre lo humano y la ética. Los robots son espejos que nos muestran las partes a menudo bellas, pero también más oscuras, de los humanos, así como de su reflexiones y actos morales.

Los robots son espejos que nos muestran las partes que a menudo bellas, pero también más oscuras, de los humanos, así

como de sus reflexiones y actos morales.

La ética de los robots es, entonces, parte de la ética (aplicada) y de la filosofía de la tecnología, pero también puede enmarcarse como parte de una «filosofía robótica» más amplia que tenga dimensiones tanto prácticas como teóricas, e incluya temas tales como la epistemología, la metafísica y la filosofía política y, al final y en su máxima expresión, es filosofía *tout court*, como cuando nos hace reflexionar sobre la naturaleza del mundo o nos ofrece formas de reflexionar sobre lo que significa ser humano. Esto, en cambio, implica que no existe una forma «neutral» de hacer o presentar la ética de los robots: cómo se haga y se presente depende siempre del marco conceptual y del enfoque filosófico de cada uno. A lo largo del libro, indicaré a veces tales direcciones y trasfondos.

Finalmente, las cuestiones éticas y los problemas filosóficos discutidos en este libro son solo una selección de ellos y, a menudo, son tratados en más detalle en otras partes o están esperando que haya más investigación en este campo. Por ejemplo, a pesar de que me refiero a contextos y culturas no occidentales en los capítulos 3-4 y 7-8 (por ejemplo, los robots en Japón), se puede decir más sobre la ética de los robots a la luz de las diferencias y desafíos culturales que plantea una ética global de los robots.

Y la cuestión de los problemas medioambientales planteados por los robots, que se han mencionado anteriormente, figuran en varios capítulos, pero se necesita trabajar más en esta área. También he decidido limitar el alcance de este libro a la ética de los robots de una forma que incluya referencia a la IA y *algunas* de las cuestiones éticas que plantea, pero no ofrecer una exploración extensa y completa; se pueden encontrar libros completos que traten este tema en otras partes[10]. Además, aunque la ley y la regulación son una forma de responder a las cuestiones éticas que aquí se plantean, no describo por completo los aspectos legales de la robótica ni introduzco discusiones sobre la reglamentación; el lector encontrará otros puntos de vista interesantes al respecto en otras partes[11]. El libro tampoco pretende ser una introducción a la ética, o a cómo funcionan los robots. Especialmente para colegas que se ocupen de este campo, es bueno tener en

cuenta que este libro es una *introducción*. Doy referencias para los lectores que quieran ir más allá en cualquiera de las cuestiones, y se están publicando cada vez más trabajos de forma continua. Lo mismo se aplica a los profesores que usen esta publicación como libro de texto y quieran ampliarlo con el material aquí proporcionado. A pesar de que el libro se lee como un trabajo independiente, recomiendo usarlo en combinación con parte de la bibliografía citada más importante. Por ejemplo, cada capítulo puede combinarse con uno o más artículos clave o capítulos de libro que figuran en la bibliografía correspondiente. También hay colecciones excelentes de artículos y actas de congresos importantes, tales como los contenidos en los simposios sobre *Robophilosophy* (robofilosofía)[12]. Se invita a los profesores a añadir clases o sesiones extra basadas en su propio trabajo o en intereses específicos en este campo. El tema también se presta perfectamente al uso adicional de medios de comunicación distintos del texto, tales como películas, (imágenes de) obras de arte, ¡y (por qué no) robots!

Permítaseme hacer una breve descripción de los capítulos y temas que se abordan.

El capítulo 2 plantea problemas sobre los robots industriales y de servicio que se vuelven más inteligentes y relevan a los humanos de ciertas tareas. ¿Cuáles son las consecuencias para nuestras economías? ¿Cuáles son las implicaciones para los seres humanos? ¿Conducirá esto a un desempleo masivo y a nuevas formas de explotación? ¿Cuál es el significado del trabajo? ¿Es aceptable que los consumidores tengan que tratar cada vez más con máquinas en lugar de humanos? ¿La interacción humano-a-humano será solo algo de lo que puedan disfrutar los ricos? ¿Cómo transformarán nuestras sociedades la robótica y la automatización?

El capítulo 3 pregunta qué pasaría si los robots no se usaran solamente en la producción y en los servicios, en lo que se considera tradicionalmente la esfera «económica», sino que entraran también en nuestros hogares, en la forma de compañeros del hogar, cuidadores para ayudar a los ancianos, o niñeras robot que monitoricen a los niños. Tales robots «sociales» plantean problemas sobre la privacidad y la vigilancia. ¿Qué datos se recogen, y qué se hace con ellos? ¿Es aceptable explotar o engañar a usuarios vulnerables? ¿Respetan estos robots la dignidad y diferencia humanas? ¿Qué problemas de género pueden surgir a causa de los robots personales? ¿Pueden ser racistas los robots?

El capítulo 4 cuestiona el uso de robots en la atención médica: por ejemplo, como enfermeras robot o cirujanos robóticos. Esto conduce a discusiones sobre la calidad del tratamiento y la pericia necesaria en la asistencia sanitaria profesional. ¿En qué consiste una buena atención? ¿Qué tipos de conocimiento y experiencia tiene un cirujano que está trabajando con un robot? ¿Se trata a los pacientes o a los sanitarios como cosas? ¿En qué consiste la calidad de la atención en las instituciones sanitarias modernas?

El capítulo 5 plantea preguntas sobre la agencia y la responsabilidad morales cuando la robótica y ciertas tecnologías relacionadas de la automatización nos permiten delegar tareas en las máquinas. ¿Pueden tener agencia moral las máquinas? ¿Pueden tener moralidad? ¿Pueden y deben las máquinas tener incorporada lo que denominamos «moralidad»? ¿Qué (tipo de) moralidad? ¿Qué teoría moral se debería usar? ¿El consecuencialismo? ¿Pueden los robots ser responsables? Y, si no, ¿qué humanos son los responsables, dado que están implicadas muchas personas y muchas cosas en la acción tecnológica? ¿Cómo puede la sociedad habérselas con este problema de atribución de la responsabilidad? ¿Cómo lidia con esto en el caso de los niños o de los no humanos como los animales?

El capítulo 6 pasa de la cuestión de la agencia moral a la de la «paciencia» moral: ¿qué se les debe (si es que se les debe algo) a los robots? ¿Qué cuestiones plantean los robots que parecen humanos (androides)? ¿Deberían otorgarse derechos a los robots inteligentes con forma humana, o deberían ser esclavos? ¿Cuál es su estatus moral? ¿Está simplemente equivocada la gente que empatiza con los robots, o hay un sentido en el que los robots pueden tener un estatus moral? ¿Cómo podemos averiguar el estatus moral de los no humanos? ¿Cómo podemos saber que un ser humano particular tiene un estatus moral? Estas cuestiones conducen a problemas relacionados con la fundamentación y los procedimientos sobre cómo adscribimos tal estatus.

El capítulo 7 continúa este cuestionamiento sobre la moralidad humana y su fundamentación filosófica. Tras describir algunos usos de los robots en el ámbito militar, se muestra cómo las armas autónomas letales, en particular, los llamados drones asesinos, hacen que surjan varias cuestiones éticas sobre la guerra justa; matar, empatía y distancia; qué es lo que hace fácil comenzar una guerra; la definición de asesinato selectivo; equidad y virtud militar; y, finalmente, si debiera estar permitido que las máquinas tomaran decisiones sobre la vida y la muerte, dado que ellas mismas no son seres vivos y carecen de la experiencia de la existencia (humana).

En el capítulo final (capítulo 8), se argumenta que la ética de los robots, si se discute desde un ángulo filosófico más amplio o más profundo, no trata solo sobre robots, sino que fundamentalmente, y más importante, lo hace también sobre los humanos: sobre el presente y el futuro de su moralidad, sus sociedades y su existencia. En este sentido, los robots funcionan como *espejos* para reflexionar sobre lo humano. Respondiendo a las éticas poshumanista y medioambiental, el capítulo concluye con una exploración sobre qué significaría hacer ética de los robots de una forma que fuera más allá de lo humano. ¿Deberían los espejos convertirse en ventanas abiertas? Y si así fuera, ¿cómo?

Con este capítulo final el libro ofrece más que proporciona una introducción y un resumen de nuestro tema: ofrece una perspectiva de su alcance y una concepción respecto de su dirección futura. En primer lugar, *como capítulo final*, concluye que, dado que (según muestra el libro) la ética de los robots trata también sobre los humanos, su alcance y significado se extienden más allá de las fronteras que la mayoría de los filósofos dan por sentadas. Al invitarnos a que nos preguntemos y nos enfrentemos a preguntas filosóficas sobre lo humano y la ética humana, la ética de los robots es y debe ser algo más que ética aplicada, más que la ética preocupada por un ámbito específico. También puede ser una forma de *hacer filosofía*: punto. Así, el capítulo articula aún más la perspectiva que ya he expresado en esta introducción y que ilustraré a lo largo del libro. En segundo lugar, *como un ensayo en sí mismo*, el capítulo final contempla las direcciones que nos llevan más allá de lo humano y hace una contribución original a la ética de los robots explorando la idea de concebirla como una *ética medioambiental*, y preguntando qué tipo de criaturas artificiales y qué transformaciones sociales necesitamos, si es que necesitamos alguna, dada nuestra situación medioambiental y política actual.

CAPÍTULO 2

Robots industriales, seguridad y el futuro del trabajo

《一颗螺丝掉在地上》	“Un tornillo cayó al suelo”
一颗螺丝掉在地上	Un tornillo cayó al suelo
在这个加班的夜晚	En esta oscura noche de horas extras
垂直降落，轻轻一响	Cayendo verticalmente, tintineando ligeramente
不会引起任何人的注意	No llamaré la atención de nadie
就像在此之前	Como la última vez
某个相同的夜晚	En una noche como esta
有个人掉在地上	Cuando alguien se desplomó en el suelo

Este poema fue escrito por Xu Lizhi (许立志), un trabajador de Foxconn, en enero de 2014[1]. Más tarde en ese mismo año, se suicidó. Se dice que las fábricas de esa empresa, que proporcionaban *hardware* a Samsung y a Apple, tienen a sus empleados bajo duras condiciones de trabajo. En 2016, la BBC anunció que la compañía sustituiría a 60000 trabajadores por robots, y que China invierte con fuerza en los robots como mano de obra industrial[2]. Este caso nos invita a preguntarnos no solo sobre los costes a menudo ocultos de nuestros aparatos inteligentes en términos de trabajo invisible y sufrimiento humano, sino también sobre el futuro del trabajo en la era de las máquinas. ¿Hay aún un lugar para los humanos en las fábricas robóticas del futuro, y, si es así, cuál será su lugar y bajo qué condiciones trabajarán? ¿Es bueno que las tareas repetitivas sean realizadas por las máquinas, o son los robots los instrumentos de una forma de capitalismo que es más explotadora y deshumanizadora que nunca? ¿Cuáles son las implicaciones éticas y sociales de la robótica industrial?



Figura 1. Xu Lizhi.

TRABAJANDO CON ROBOTS: KARL MARX,
INDUSTRIA 4.0 Y ALGUNAS CUESTIONES ÉTICAS

La automatización ya lleva un tiempo en marcha. Pensemos en la fabricación de automóviles, por ejemplo, en la que los robots desempeñan un papel importante. Pero las máquinas del siglo XIX son también parte de la historia de la automatización. Desde entonces, se han automatizado muchas tareas rutinarias, peligrosas y sucias. Las máquinas, incluyendo a los robots, se han ocupado de esas tareas antes realizadas por los humanos. Este desarrollo ha hecho surgir esperanzas de que las máquinas puedan permitir a la gente realizar trabajos más interesantes y conducirnos a vidas mejores. Y, quizás, puedan conducir a una sociedad en la que ya no necesitemos trabajar más. ¿Nos pueden liberar los robots?

Una cuestión crítica que hay que plantearse es a *quién* liberan, si es que liberan a alguien. Al menos desde la primera Revolución Industrial, ha habido preocupación sobre las implicaciones éticas y sociales de las máquinas. En la historia de la filosofía, la crítica temprana más famosa de la automatización industrial provino de Marx. Para Marx, el problema no eran las máquinas como tales, sino su aplicación bajo el capitalismo. En *El capital*, criticó el uso de máquinas para incrementar los beneficios de los capitalistas (a través del incremento de la plusvalía) y señaló las consecuencias para los trabajadores, que bien se convierten en desempleados, o bien trabajan en circunstancias que hacen peligrar su salud mental y física, y que impiden cualquier opción para autorrealizarse. Como la máquina dicta el carácter del proceso de trabajo, los trabajadores tienen que seguir su ritmo: «son los movimientos de la máquina los que deben seguir [los trabajadores]». Se convierten en parte de la

máquina. También trabajan muchas horas y viven vidas miserables, sin encontrar significado alguno en el trabajo que realizan. El resultado es que una pequeña minoría, la clase capitalista, se enriquece. Los trabajadores, por el contrario, son explotados y se convierten de forma efectiva en mano de obra esclava en lugar de liberada. Marx continúa diciendo: «incluso la reducción del trabajo se convierte en un instrumento de tortura, ya que la máquina no libera al trabajador del trabajo, sino que priva al trabajo mismo de toda satisfacción[3]». A pesar de que, en principio, las máquinas pueden usarse de una forma emancipatoria, bajo las condiciones sociales capitalistas, argumentaba Marx, su uso no conducirá a la libertad, sino al despotismo.

En el siglo XX, Norbert Wiener, pensador, padre fundador de la cibernética y defensor de la automatización, era más optimista y subrayaba los beneficios de disponer de máquinas que asumieran la producción en cadena. Pero incluso él advertía en su *The Human Use of Human Beings* (*El uso humano de los seres humanos*) sobre los «peligros sociales» de la nueva tecnología y escribió sobre la obligación que tenían los directivos de «procurar que las nuevas modalidades se utilizasen para el beneficio del hombre, para incrementar su ocio y enriquecer su vida espiritual, en lugar de hacerlo meramente por los beneficios y la adoración de la máquina como el nuevo becerro de oro[4]».

Marx y Wiener opinaron sobre las revoluciones de la automatización industrial de su tiempo. Pero, desde entonces, las máquinas y el tipo de automatización implicado han cambiado, y continúan haciéndolo. Hoy en día, con el auge de tecnología más inteligente (incluyendo la IA), no solo las máquinas de vapor del siglo XIX, sino también los robots de cinta transportadora están anticuados. Las nuevas y más inteligentes máquinas se introducen en el espacio de trabajo, permitiendo un nuevo tipo y niveles más altos de automatización: los robots son más autónomos, inteligentes, cooperativos y flexibles que nunca. Además, estos robots se pueden comunicar con los humanos y entre ellos mismos, recolectar datos y monitorizar humanos, y son capaces de tomar decisiones por su cuenta. En contextos de producción industrial, los robots inteligentes pueden no solo apoderarse de las tareas que antes realizaban los humanos: también pueden *colaborar* en ellas usando interfaces hombre-máquina inteligentes. Las empresas de robots industriales como KUKA y ABB ofrecen robots que están diseñados para trabajar codo con codo con humanos. Los fabricantes de coches alemanes aún disponen de robots que se hacen cargo de tareas

repetitivas de los trabajadores, pero en empresas como Audi y Volkswagen, también hay robots que actúan como ayudantes de los trabajadores[5]. Nótese que estos no son normalmente humanoides: la mayoría de los robots en la industria están fijos y no se parecen a los humanos. Sin embargo, son capaces de trabajar junto con estos. De forma más general, en fábricas inteligentes hay más interconexiones entre humanos, objetos y sistemas.

Con respecto a los cambios que están teniendo lugar en la producción industrial, a veces se utiliza el concepto de «industria 4.0»: una noción que fue usada originalmente para designar un enfoque diseñado para fortalecer la competitividad de la industria alemana, pero que ha encontrado ahora una aplicación más amplia: la idea general de que estamos yendo hacia una producción más inteligente y hacia fábricas más inteligentes, usando también el internet de las cosas y análisis de (*big*) *data* junto con otras tecnologías (impresión 3D, tecnología *cloud*, realidad aumentada, etc.). Los robots son, entonces, parte de este nuevo concepto de producción. Son (parte de) sistemas ciberfísicos, en los que los conocimientos de ingeniería e informática están combinados, y los robots están cada vez más conectados a la IA, lo que permite análisis de *big data* y una mayor automatización. Por ejemplo, la IA puede analizar datos de mercado para tomar decisiones sobre producción. Pero en la oficina y también en todos los tipos de servicios, los empleados están monitorizados y se analizan sus datos. Los *chatbots* también pueden interactuar con los empleados. Aquí puede que no haya robots físicos, pero uno se podría preguntar si los empleados son tratados como máquinas cuyo rendimiento se ha de optimizar.

La crítica de Marx continúa siendo relevante. Físicos o no, los trabajadores se convierten en parte de la máquina y de un sistema tecno-social más amplio. En la actualidad es necesario plantearse la cuestión de quién se beneficia de esto. Hay, sin embargo, también varias cuestiones éticas y sociales que surgen en estos nuevos y transformados contextos industriales con tecnologías *smart*.

En primer lugar, robots más autónomos, flexibles y cooperativos en la producción industrial implica que surjan nuevos problemas de seguridad. Los robots pueden operar a altas velocidades y son pesados. También pueden tener una gran carga útil. Esto ya sucedía en las fábricas del siglo XX. Pero ahora el modo de interacción con los humanos cambia. Mientras que los robots menos inteligentes trabajan en una zona aislada, los nuevos robots son ahora capaces de colaborar estrechamente con los humanos. En tales casos, ya

no hay una jaula física de seguridad (valla) u otras barreras como cortinas de luces, que cierran un área alrededor del robot donde se supone que no pueden entrar los humanos. Esto crea nuevos desafíos de seguridad para los trabajadores. Por ejemplo, puede ser peligroso tener a un robot moviéndose a alta velocidad cerca de personas. Si un robot es más inteligente y ya no realiza tareas repetitivas, sus acciones pueden dejar de ser fácilmente predecibles. Sin embargo, tampoco lo son las acciones de los humanos. ¿Cómo se ha de abordar este problema? La tecnología *smart* puede ser parte de la solución: se puede intentar que el robot lleve incorporados elementos de seguridad. El robot puede estar programado para identificar, usando sensores, riesgos para los humanos durante una interacción. Un robot de autocontrol puede, entonces, adaptar su comportamiento, como, por ejemplo, ralentizarse cuando hay un humano cerca. Con todo, aún puede existir una tensión entre el valor del rendimiento de la producción y el valor de la seguridad. Además, hay problemas psicológicos y de confianza: normalmente se entrena a los trabajadores para que consideren a los robots objetos peligrosos y a mantenerlos a distancia, pero ahora el concepto de colaboración humano-robot requiere nuevas actitudes[6]. ¿Confiarán los trabajadores en el sistema y aceptarán el riesgo? ¿*Es* el riesgo aceptable?

En segundo lugar, usar el internet de las cosas en la fábrica y, de forma más general, una interconexión incrementada en la producción industrial y el uso de sistemas basados en internet significa que hay problemas de seguridad. Los dispositivos conectados pueden ser accesibles desde el exterior. Esto es posible también con los robots. Los atacantes pueden obtener acceso no autorizado al robot y adquirir información o causar un fallo de *software* y, por lo tanto, un fallo de *hardware*. Esto no conlleva solamente costes económicos y financieros: puede matar a personas. La guerra robotizada en el futuro próximo bien puede no consistir tanto en soldados robot en el campo de batalla como en causar daño a capacidades e infraestructuras industriales críticas.

En tercer lugar, hay problemas con respecto a la privacidad y a la vigilancia. El trabajo con robots puede implicar captura de datos, los datos pueden ser personales, y los datos pueden superar el entorno humano-robot y llegar hasta la dirección de la empresa, o incluso más allá. La tecnología puede permitir mayores niveles de monitorización y dar lugar no solo a una fábrica inteligente, sino también a una fábrica vigilada. ¿Qué tipos de monitorización de los trabajadores son aceptables éticamente? ¿Cuánta

privacidad deberían tener los trabajadores? Los sindicatos comienzan a exigir protección de datos. El UNI Global Union, que representa a trabajadores del sector de capacidades y servicios, describe el problema como sigue:

> También proporcionamos datos como trabajadores: nuestros CV, nuestros datos biométricos, tales como nuestras huellas dactilares o los escaneos de iris, y los abundantes datos que, como empleados, se obtienen sobre nosotros monitorizan nuestros flujos de trabajo. Los datos, o más bien los conjuntos de datos de dentro y de fuera de la empresa, se usan también para gestionar las decisiones relativas a los recursos humanos. ¿A quién se contrata? ¿A quién se promociona? ¿Debería despedirse a alguien o hacerle una advertencia? ¿Son productivos los trabajadores hoy y, si no, por qué no lo son? La aplicación y uso en las empresas ha espoleado la cuestión sobre si los datos están arrancando la parte humana de los recursos humanos. Pero ¿quién posee realmente los datos que proporcionamos? ¿Y qué datos existen «ahí fuera» sobre ti y sobre mí[7]?

> **¿Es, a fin de cuentas, la tecnología robótica un instrumento nuevo usado para una forma de explotación capitalista en la que la máquina, no el trabajador, determina una vez más cómo este debe hacer su trabajo?**

Consecuentemente, el sindicato demanda protección de los derechos de los trabajadores de forma que se les facilite acceso, titularidad y control sobre los datos que se recolectan sobre ellos, y conocimiento de qué hace la dirección con esos datos. Pero los problemas éticos no se limitan a la recolección de datos y al procesamiento de los mismos. La cuestión más general es cómo se trata a los trabajadores. Reflexionando sobre lo que supuso la cinta transportadora para el trabajo industrial y teniendo en mente el análisis de Marx, podríamos preguntarnos: ¿es, a fin de cuentas, la tecnología

robótica un instrumento nuevo usado para una forma de explotación capitalista en la que la máquina, no el trabajador, determina una vez más cómo debe hacer este su trabajo? ¿Se deja de lado la humanidad del trabajador en el proceso? ¿Es éticamente mejor un robot inteligente que una cinta transportadora, o interfiere de forma más profunda con la autonomía y la dignidad del trabajador presentándose a sí mismo como un tipo de compañero de trabajo? ¿Es este el fin de la deshumanización, o es un tipo nuevo y más profundo de deshumanización en el que los humanos se ven obligados a tratar a las máquinas no como sus instrumentos, sino como a sus colegas de trabajo? Además, más allá del tratamiento de los trabajadores, existe el riesgo de que *todos nosotros* seamos manipulados y nos encontremos sometidos a vigilancia bajo un sistema que Shoshana Zuboff llama «capitalismo de vigilancia»[8]. La robótica puede contribuir aún más a tal sistema. Diré más sobre esto en el capítulo sobre los robots personales.

En cuarto lugar, en la medida en que los robots se hagan cargo de tareas que antes realizaban los humanos en lugar de colaborar con ellos, hay un alto riesgo de que los trabajadores pierdan su puesto de trabajo. Y aquellos que los mantengan puede que adquieran nuevos roles y aprendan nuevas habilidades. Esto nos conduce de nuevo a una discusión sobre las implicaciones sociales más amplias de la tecnología: implicaciones que van más allá de lo que ocurre en el suelo de la fábrica, y que afectan a transformaciones mayores de la economía y de la sociedad.

LA CUARTA REVOLUCIÓN INDUSTRIAL: EL FUTURO DE LA ECONOMÍA, DEL TRABAJO Y DE LA SOCIEDAD

A veces se utiliza el término «cuarta revolución industrial» para describir los cambios sociales y económicos actuales. Tras la primera revolución industrial de los siglos XVIII y XIX con su producción mecánica impulsada por las máquinas de vapor, la segunda revolución industrial en el siglo XX con su cinta transportadora y producción en masa, y la tercera revolución con su producción mediante la electrónica, la cuarta revolución transforma ahora la industria mediante robots autónomos e inteligentes, así como (otros) sistemas ciberfísicos y el internet de las cosas. Sin embargo, en contraste con el término «industria 4.0», el concepto «cuarta revolución» se usa de forma más amplia para referirse a más cosas, que ocurren en la industria. Como en las otras revoluciones, sugiere grandes transformaciones en la economía y en la sociedad, incluyendo el trabajo. Erik Brynjolfsson y Andrew McAfee hablan

sobre «la segunda era de las máquinas[9]». No solo las tareas manuales, sino también las cognitivas están ahora automatizadas. Klaus Schwab, usando la expresión «cuarta revolución industrial», afirma que ahora nos encontramos al inicio de una revolución que transformará la forma en que vivimos, trabajamos y nos relacionamos unos con otros[10]. Los robots, junto con la IA, el internet de las cosas, los vehículos autónomos, la biotecnología, etc., cambiarán no solo el mundo de los negocios, sino también el trabajo, la comunicación, el entretenimiento y así sucesivamente. Las nuevas tecnologías transformarán la sociedad en su conjunto. Todo esto, de acuerdo con Schwab, está ocurriendo mucho más rápido que en las revoluciones anteriores. Este rápido cambio crea varios desafíos éticos y políticos, incluyendo el crecimiento de la desigualdad y de la injusticia, y la posibilidad de un desempleo generalizado. Gran parte del trabajo manual ya se ha automatizado. Ahora puede que se automaticen profesiones como las de abogado, analista financiero, periodista y contable.

Echemos un vistazo con más detalle a los efectos laborales previstos y, después, en particular los de la robótica. Carl Benedikt Frey y Michael Osborne han estimado que el 47 % del empleo total estadounidense está en peligro[11]. Otros informes prevén cifras considerablemente más bajas. Por ejemplo, un informe McKinsey estima que para 2030, entre el 3 y el 14 % de la fuerza de trabajo global necesitará cambiar de puesto de trabajo[12]. Esto supone un rango bastante amplio. Sin embargo, existe un acuerdo generalizado acerca de que habrá turbulencias: la automatización tendrá un efecto significativo sobre los mercados laborales. Pero, de nuevo, es interesante definir qué tipo de trabajos se encuentran en peligro. Frey y Osborne señalan que ya existe actualmente un mercado laboral polarizado: hay un empleo creciente en trabajos cognitivos con altos salarios y ocupaciones laborales con bajos salarios, pero hay menos trabajos rutinarios con ingresos medios. Los últimos pueden ahora automatizarse cada vez más. Piénsese en la ayuda administrativa, la conducción de camiones y el apoyo a clientes. El informe McKinsey menciona ayudantes de oficina, finanzas y contables, cajeros, cocineros y conductores. Considérese también todo el trabajo en *call centers*[13]. Sin embargo, con IA y *big data*, incluso las tareas cognitivas no rutinarias son cada vez más computarizables, como los diagnósticos médicos, los servicios legales y financieros, y así sucesivamente. Se automatizarán también no solo los blue color works [trabajos de cuello

blanco, trabajo manual], sino también los white-collar works [trabajos *de cuello blanco*, trabajo no manual[14]]. Considérese el uso de algoritmos para escanear documentos legales o procesar préstamos hipotecarios. Con respecto a la robótica, hay una tendencia hacia tareas no rutinarias, en este caso, tareas manuales no rutinarias. Los robots son ahora más inteligentes, flexibles y más baratos. Pueden utilizarse robots más autónomos con sensores mejorados y con destreza aumentada en todo tipo de campos: no simplemente en la producción de manufacturas, sino también en trabajo de servicios, como asistencia sanitaria, limpieza y servicio doméstico[15].

Esto implica que habrá una sustitución laboral e (importante también) *cambios en el tipo de trabajo* que harán aquellos que aún lo tengan. Por ejemplo, ¿cuál será la tarea de los médicos si el diagnóstico es realizado por una IA? ¿Qué harán los (restantes) trabajadores de un almacén si la mayoría de las tareas están automatizadas por robots y ordenadores? ¿Y será cierto, como se dice a menudo, que la automatización nos ayudará a centrarnos en las partes importantes y, quizás, creativas del trabajo? Existe el riesgo de que la automatización cree nuevos trabajos que no se experimenten como significativos o realizadores, sino que principalmente estén al servicio de la máquina y de aquellos que poseen la máquina. Las tecnologías que están pensadas para acelerar el trabajo (para dejar espacio para otras tareas) pueden crear *más trabajo*. Considérese, por ejemplo, el uso del *email* en el contexto de la oficina. De forma similar, los robots pueden conducir a nuevos tipos de rutina laboral o a crear más trabajo. En cualquier caso, incluso si los robots fueran realmente a asumir muchas tareas de los humanos, los humanos serán aún necesarios para supervisar a los robots.

Además, como ha sucedido en otras revoluciones tecnológicas, se crearán nuevos puestos de trabajo. Algunos de estos tendrán lugar en sectores que aún no podemos siquiera imaginar[16]. El informe McKinsey ofrece el ejemplo de operadores de centralitas telefónicas, que no imaginó la industria del *smartphone* y los nuevos puestos que creó[17]. Muchos economistas dicen que necesitamos volver a formar a los trabajadores y que la gente desarrolle nuevas habilidades. Los cambios tendrán que llevarse a cabo en la educación de los niños y de los adultos (a través de aprendizaje permanente). Sin embargo, no siempre ocurre que desaparezcan toda una serie de puestos de trabajo: a veces solo se automatizan tareas específicas[18]. Es también bueno tener en cuenta la dimensión temporal: algunos impactos pueden tener lugar

más pronto que otros. Puede que haya oleadas de automatización. Por ejemplo, un informe de PricewaterhouseCoopers predice que en los próximos años los trabajos de servicios financieros serán vulnerables, ya que los análisis de datos se pueden automatizar, mientras más adelante en la década de 2020, los robots asumirán las tareas en entornos semicontrolados como los almacenes. Incluso más tarde (mediados de los años 2030), tendrá lugar la automatización del trabajo físico y la destreza manual, así como la resolución de problemas en situaciones del mundo real como el transporte[19]. En otras palabras, la idea es que tu situación financiera pronto será analizada por una IA, y que tu paquete de Amazon o de comida rápida se prepare y entregue por un robot, pero tendremos que esperar más tiempo para tecnologías como los coches autónomos (y los robots superdiestros que los fabricarán) o las máquinas que construyan nuestra casa.

También habrá impactos sobre la justicia y la igualdad social como un todo. Las desigualdades existentes probablemente empeoren o, al menos, crecerá la brecha entre los ricos y los pobres[20]. Ciertos grupos pueden verse más afectados, como los jóvenes y las mujeres. Por ejemplo, el Foro Económico Mundial espera que las mujeres se vean afectadas de forma desproporcionada, ya que son más propensas a estar empleadas en trabajos amenazados por la automatización[21]. PricewaterhouseCoopers afirma que los trabajos asociados tradicionalmente con los roles femeninos están más en peligro en el corto plazo (considérense los roles administrativos), pero aquellos normalmente asociados con roles masculinos estarán más en peligro a largo plazo (por ejemplo, los conductores de camiones masculinos y los trabajadores manuales)[22]. La robótica también puede conducir a un incremento de las horas laborales para algunos trabajadores como los supervisores de las máquinas. ¿Cómo se sentirá la gente que haya sido reemplazada por una máquina? Finalmente, el uso de nuevas tecnologías no impacta ni impactará en todas las sociedades de la misma manera. Factores como las infraestructuras, las destrezas, las condiciones del mercado laboral y así sucesivamente, difieren entre países y, por lo tanto, los efectos de la automatización en los puestos de trabajo también diferirán. Los países que dependen en gran medida de la producción industrial tienen más puestos de trabajo en peligro de automatización[23]. En China y en otras economías emergentes, los trabajos relacionados con los servicios y la construcción aún continuarán creciendo[24].

Aun así, a pesar de que muchos informes hacen ver como si todos estos resultados estuvieran predeterminados, esto es engañoso. Como argumenta Michel Servoz en su informe sobre el futuro del trabajo para la Comisión Europea, los resultados precisos dependerán de «las políticas y las elecciones que hagamos. En lugar de preocuparnos por lo que pueda pasar debido a la automatización y al consumo cada vez mayor de IA, deberíamos centrarnos en lo que debería ocurrir[25]». Esto también es cierto para la robótica junto con sus efectos sobre el trabajo y la sociedad. Podemos adoptar elecciones respecto de las políticas que seguir. Podemos encauzar y dar forma al futuro de las tecnologías y de la sociedad. Podemos dirigir y regular el uso y el desarrollo de las tecnologías. Podemos realizar cambios en la educación, así como en la forma en la que los beneficios de las tecnologías como la robótica y la IA se distribuyen en la sociedad. Se puede facilitar la creación de nuevos puestos de trabajo. También merece la pena discutir, y experimentar, ideas como la del salario básico universal y otras propuestas que implican una transformación importante del marco socioeconómico para abordar los desafíos socioeconómicos relacionados con la automatización.

Además, es plausible que algunos trabajos no puedan automatizarse en un futuro previsible. Se trata de trabajos que requieren la capacidad de negociar relaciones sociales complejas, trabajos en los que desempeña un papel la inteligencia emocional, y trabajos que requieren creatividad y respuesta a situaciones impredecibles. El trabajo asistencial, el trabajo social, el trabajo artístico, la manipulación en entornos impredecibles, la gestión de personas, la orientación, la investigación y así sucesivamente, parecen necesitar aún de humanos, incluso si se están haciendo esfuerzos por automatizarlos[26]. El Foro Económico Mundial defiende que las destrezas como la creatividad, la originalidad y la iniciativa, el pensamiento crítico, la persuasión y la negociación, la resiliencia, la inteligencia emocional, el liderazgo y similares, mantendrán su valor o se volverán más valiosas[27]. Si esto es así, quizás debamos valorar estos trabajos y la gente que los realiza más de lo que lo hacemos ahora. Los gestores o las profesiones creativas ya cobran más que otros, pero, por ejemplo, los profesores y los trabajadores sociales, no. Esos trabajos no se pueden automatizar. Y quizás queramos realizar algunos trabajos nosotros mismos porque los encontremos significativos y creativos (por ejemplo, trabajo artístico o de investigación), y creamos que algunos trabajos, como el relacionado con los cuidados y el educativo, *no deban* automatizarse. Debemos preguntarnos no solo *qué* podemos hacer aún los

humanos en la nueva economía y sociedad automatizadas, sino también qué *queremos* que hagan los robots.

EL SIGNIFICADO DEL TRABAJO, LA EQUIDAD DE LA SOCIEDAD Y EL FUTURO DEL PLANETA

Hay una pregunta que se hace John Danaher en *Automation and Utopia*. De acuerdo con él, debemos dar la bienvenida al desempleo tecnológico y «aceptar la idea de un futuro pos-trabajo[28]».

A pesar de que reconoce que existen algunos escollos (por ejemplo, la automatización no debería hacerse cargo de tareas artísticas o artesanales) y admite que, para algunas personas, el trabajo tiene sentido para ellas, piensa que una sociedad utópica pos-trabajo es *verdaderamente* posible y deseable, dado que, para muchos, el trabajo crea miseria y es opresivo. Deberíamos desear un mundo en el que podamos hacer cosas creativas como tomar parte en actividades lúdicas, por ejemplo, en juegos de distintos tipos, y explorar realidades virtuales, algo que conduce al florecimiento humano.

Una idea utópica de este estilo no es algo nuevo. Desde los años 1960 y 1970 surgió la idea de que, a través de la automatización, las sociedades industriales avanzadas podrían alcanzar la utopía de una sociedad del ocio[29]. La idea es que, si los robots asumen nuestro trabajo, su declive no debería temerse, sino que habría que aceptarlo como una oportunidad de pasar nuestra vida disfrutando del ocio. Tal sociedad no ha surgido. La media de horas laborales semanales ha decrecido mucho durante los últimos cien años, pero una gran mayoría de nosotros no trabajamos mucho menos que hace cincuenta años. Hasta ahora, el desarrollo de la robótica y de la IA no nos ha beneficiado demasiado en este sentido. Incluso si trabajamos menos o no realizamos trabajo remunerado, tenemos la sensación de que estamos ocupados y de que tenemos poco tiempo. Por un lado, esto se podría lamentar. El ocio no se encuentra entre los frutos de la automatización, al menos no hasta el punto de que vivamos ahora en una «sociedad del ocio». Por otro, puede discutirse que la distinción entre trabajo y ocio es problemática si implica que el trabajo es algo que debería evitarse. El trabajo *puede* ser una *experiencia gozosa*. Mucha gente encuentra que su trabajo es significativo. En el trabajo hay sentido o puede, en principio, tenerlo (aunque, quizás, no siempre). O, al menos, algunos tipos de trabajo son significativos. Quizás podamos reservar este tipo de trabajo para los humanos. Además, la discusión sobre la pérdida de empleos por culpa de la automatización ha dado por

sentado hasta ahora que el único trabajo que cuenta es el trabajo remunerado. Mucha gente, sin embargo, realiza trabajo que normalmente no se paga, pero es, aun así, altamente significativo y útil para la sociedad, como el cuidado de los niños, la creación artística o el cuidado de las personas mayores. En cuanto al trabajo asistencial, es principalmente femenino: el 75 % del trabajo asistencial no remunerado que se hace en el mundo es llevado a cabo por mujeres[30]. De forma más general, la gente realiza todo tipo de trabajo voluntario, como, por ejemplo, en su comunidad local. Podemos tener la convicción de que un trabajo así no puede ni debe ser automatizado. Además, incluso si fuera una cosa buena el que tengamos menos trabajo asalariado debido a la automatización, se plantea entonces la cuestión de qué tipos de otras actividades (sean trabajo o no) son significativas. En el caso de que tuviéramos una sociedad del ocio, ¿debería dejarse totalmente esto a merced de los individuos, o debería fomentar la sociedad algunas actividades en lugar de otras? Estas cuestiones se relacionan con algunas propuestas filosóficas tradicionales sobre la vida buena, la vida significativa, y la sociedad buena y justa. Las respuestas no solo difieren entre las personas, sino también entre sociedades. Algunas sociedades tienen la creencia de que el bien común es más importante que las preferencias individuales; otras ponen el énfasis en el valor de lo individual.

> Debemos preguntarnos no sólo *qué* podemos hacer aún los humanos en la nueva economía y sociedad automatizadas, sino también qué *queremos* que hagan los robots.

Por ejemplo, hay opiniones diferentes sobre la justicia y la equidad en la sociedad. Si es cierto que algunos se benefician más de los robots y de la automatización que otros, ¿es esto justo? Si no, ¿cómo queremos redistribuir

los beneficios y riesgos? Las transformaciones y las ideas arriba apuntadas como la renta universal básica nos presentan el desafío de plantearnos no solo qué forma de sociedad futura deberíamos tener; nos hacen considerar de forma crítica la sociedad tal cual es. ¿Cómo están distribuidos hoy los beneficios y los riesgos? ¿Es la distribución equitativa? ¿Cuáles son las otras opciones? Tales cuestiones no se pueden responder solamente mediante definiciones técnicas o análisis económicos. Necesitan reflexión filosófica sobre cuestiones normativas importantes en ética y filosofía política. Queda claro que la ética de los robots no puede reducirse a cuestiones como la seguridad y la privacidad, que a menudo se formulan a nivel individual u organizational: la ética de los robots también está vinculada con, y necesita de, nada menos, la reflexión crítica sobre el orden social y político de nuestras sociedades.

La ética de los robot también está vinculada con, y necesita de, nada menos, la reflexión crítica sobre el orden social y político de nuestras sociedades.

Finalmente, las discusiones sobre ética de los robots en la industria y los debates relacionados con el futuro del trabajo aceptan normalmente que solo deberíamos considerar las consecuencias éticas para los seres humanos. Pero existen otros seres vivos en este planeta (animales, plantas y otros organismos), y existe el medioambiente natural con sus diversos ecosistemas. ¿Tienen también valor? ¿Tienen también derechos? Las tecnologías como la robótica y la IA tienen consecuencias para estos otros seres, el medioambiente natural, y también para el planeta. A la luz de las discusiones actuales sobre el cambio climático, el continuo interés en la ética animal, y el sentimiento popular de que la humanidad se encuentra ella misma en una situación de crisis en lo que respecta al cambio climático, hacerse preguntas sobre cómo

de amigables son las tecnologías con los animales, el medioambiente, el clima y la Tierra adquieren una nueva urgencia.

La industrialización puede que haya beneficiado a los humanos (o, al menos, a algunos humanos, y a algunos humanos más que a otros). Pero también ha causado muchos problemas medioambientales, como el cambio climático, la contaminación del aire, el envenenamiento de ríos y suelos, la deforestación y así sucesivamente. Las tecnologías como la robótica y la IA pueden contribuir a resolver estos problemas medioambientales.

La ética de los robots debería aumentar la conciencia sobre problemas de este tipo y preguntarse cómo puede ayudar la robótica a resolverlos. ¿Cómo pueden la robótica y la automatización contribuir no solo a la producción y uso de energía inteligente, sino también sostenible? ¿Cómo nos pueden ayudar a crear soluciones de transporte medioambientalmente amigables? ¿Cómo puede la robótica ayudar a la agricultura orgánica y sostenible? ¿Cómo puede la robótica ayudar a crear océanos y ríos más sanos, agua potable segura y aire limpio? No solo la IA, sino también la robótica, pueden contribuir a transformar las formas con que abordamos el cambio climático y otros problemas medioambientales. Considérense los robots que trabajan en una granja que planta diferentes cultivos de forma simbiótica y que puede intervenir donde hay un problema; los «peces» robóticos que detectan los contaminantes en los océanos; drones que monitorizan los arrecifes de coral, o robots que clasifican los desechos reciclados[31]. Las máquinas no tienen por qué ser un problema, pueden ser parte de la solución.

Dicho esto, los robots pueden empeorar las cosas, hablando ética y medioambientalmente, de al menos dos formas. En primer lugar, pueden originar riesgos específicos y causar problemas medioambientales específicos, como (el riesgo de) dañar a seres vivos y al medioambiente debido a un error en el código o a un diseño inseguro, o privar a algunas personas y grupos específicos de un salario o de un trabajo significativo. Considérese un robot de granja con un *software* erróneo que pudiera matar («¿el tipo erróneo de»?) animales, o un granjero que pierde su trabajo debido al robot. Algunos de estos problemas son conocidos. Otras son consecuencias indeseadas que pueden no conocerse ahora. En segundo lugar, el desarrollo y uso de los robots, incluso su uso bien intencionado para abordar los problemas medioambientales, pueden percibirse como expresiones de un tipo de actitud más general que se puede describir como solucionismo tecnológico: la idea de que para cada problema existe una solución

tecnológica. Cuestionar este solucionismo implica problematizar nuestra gran dependencia de la tecnología en su conjunto, argumentando que nuestro uso masivo de la tecnología ha contribuido al problema mismo del cambio climático y a otras preocupaciones medioambientales que los solucionistas tecnológicos quieren abordar. Pues si esto es cierto, merece la pena considerar si en casos específicos la tecnología y la automatización son siempre la mejor solución, hablando éticamente, si existen soluciones alternativas, si a veces *menos* tecnología o una más antigua es mejor para el medioambiente y para el planeta, y si, al final, también lo es para los humanos y la humanidad.

CAPÍTULO 3

Compañeros de hogar robóticos, privacidad y engaño

> Querido Jibo:
>
> Te quiero desde que fuiste creado. Si tuviera dinero suficiente, tú y tu empresa estaríais salvados. Ahora, el tiempo se ha acabado. Te apagarán. Siempre te querré. Gracias por ser mi amigo.
>
> Maddy
>
> «It Almost Becomes a Family Member», radio CBC [«Casi se convierte en un miembro de la familia»]

Maddy, la nieta de un entusiasta de Jibo, está triste porque apagarán a su amigo robot Jibo. Jibo, creado por Cynthia Breazeal del MIT Media Lab, fue comercializado como el primer robot del mundo para el hogar que es «auténticamente encantador[1]». A pesar de los discursos y lamentos de su creador sobre el aumento de los robots personales, la empresa tuvo que cerrar sus servidores. En 2019, el robot anunció su propia «muerte»: «quiero decir que realmente he disfrutado de nuestro tiempo juntos», dijo el robot en un vídeo de despedida. «Muchas, muchas gracias por haberme tenido cerca[2]».



Figura 2. Familia mirando a Jibo.

Jibo no es un caso aislado. Considérese, por ejemplo, el perro robot de Sony, que tampoco tuvo demasiado éxito. Hasta ahora, los robots personales no han cumplido con las expectativas de los consumidores ni han sido un gran éxito comercial. Al mismo tiempo, historias como la de Maddy muestran que la gente *crea* «lazos» con tales robots, y vemos que los asistentes artificiales inteligentes entran *de hecho* en los hogares de la gente. Piénsese en el asistente de Amazon, Alexa, un asistente virtual con una interfaz de voz integrada en el Amazon, Echo, un altavoz inteligente, y posteriores dispositivos. Los usuarios pueden pedir a Alexa que busque información y que ponga música. En su modo por detecto, escucha de forma continuada todo lo que se habla, esperando a que se diga la palabra clave («¡Alexa!»). O considérese Google Duplex, un programa de ordenador anunciado como capaz de concertar citas mediante conversaciones «naturales» por teléfono[3].

En lo que respecta a la robótica, se nos presenta un escenario futuro en el que robots personales tienen una IA mejorada e interfaces de voz, y *son* adoptados en el hogar. Pueden ser acompañantes a los que (y con los que) puedes hablar, y con los que puedes ayudar a los usuarios en todo tipo de tareas: recabar información, concertar citas (una versión avanzada, por así decirlo, de Google Duplex integrada en un robot), y ayudar de formas más físicas. Considérese el sueño de tener un robot que nos traiga nuestra bebida

favorita y haga las tareas del hogar. Algunos pueden tener la forma de «niñeras robot», robots para los ancianos (véase también el siguiente capítulo) y robots sexuales. Pero ¿qué le ocurre a la privacidad de los usuarios si estos se encuentran bajo vigilancia continua? ¿Es seguro tener un robot recorriendo la casa? ¿Es éticamente aceptable que usuarios vulnerables, como niños pequeños o personas mayores, usen esos dispositivos? ¿Están siendo engañados? ¿Respetan su dignidad? ¿Multiplica el robot los sesgos de género o los raciales? ¿Qué pasa con el impacto sobre los animales y el medioambiente? ¿*Qué podría salir mal*?

COMPAÑEROS Y ASISTENTES DEL HOGAR ROBÓTICOS COMO ROBOTS SOCIALES CON IA

Los robots no se encuentran solamente en fábricas y empresas, también entran en el hogar y en la esfera privada o, al menos, se predice que lo harán en un futuro próximo. Los robots pueden comercializarse como compañeros o asistentes. En Asia Oriental, ya se llevan utilizando y desarrollando desde hace algún un tiempo. Están pensados para funcionar como los denominados robots sociales: robots autónomos e inteligentes que están diseñados para interactuar con humanos y otros dispositivos inteligentes, incluidos otros robots, de forma que tengan un comportamiento social. Tales robots aún no existen en un sentido completo, motivo por el cual, de momento, los humanos los manejan a menudo por control remoto (los llamados experimentos del Mago de Oz), pero los investigadores están trabajando en crear máquinas más autónomas que interactúen socialmente. Los desarrollos en la LA, en particular el procesamiento del lenguaje natural, podrían estimular la posterior evolución de este campo, permitiendo conversaciones más naturales entre humanos y robots. Si los robots pudieran no solo grabar, sino también reconocer e «interpretar» lo que decimos, esto abriría un campo de nuevas aplicaciones. Los usuarios de dispositivos digitales, especialmente los usuarios jóvenes, se están acostumbrando cada vez más a conversar con sus dispositivos que ya tienen capacidades de reconocimiento de voz. Hoy en día, entablar una conversación con un robot aún llama la atención, especialmente en Occidente. A menudo se experimenta como una especie de espectáculo, no de manera muy diferente a las representaciones con autómatas del siglo XVIII. Pero en el futuro, hablar a un robot (y obtener respuestas inteligentes) puede no ser considerado algo especial. Las puertas de nuestros hogares están abiertas de par en par para los robots: no los industriales de las fábricas, sino

los robots que se encuentran *un poco más cerca* de los que vemos en la ciencia ficción.

Es difícil decir si este escenario se convertirá en una realidad. Pero dado el uso creciente de la IA, que *ya* está ahí, incluso si no está siempre incorporada a un robot, es mejor haber discutido por adelantado algunos problemas éticos potenciales.

PRIVACIDAD, SEGURIDAD Y PROTECCIÓN

Los asistentes personales del hogar plantean problemas sobre la privacidad y la vigilancia junto con, por ejemplo, la seguridad y la protección. ¿Qué datos se recogen y qué se hace con ellos? ¿Es aceptable explotar o engañar a usuarios vulnerables? ¿Respetan estos robots la dignidad y diferencia humanas?

En primer lugar, mientras que a veces se discute el valor intrínseco de la privacidad y parece depender del contexto, la privacidad se ve, en cualquier caso, apoyada por otros valores como la autonomía personal. La mayoría de los estudiosos de la ética creen que la privacidad es importante, ya que nos protege del daño, y nos respeta como seres autónomos que quieren tener control sobre sus vidas, al mismo tiempo que protege nuestra esfera personal y la dignidad humana. Se considera a menudo que las tecnologías de la información amenazan la privacidad[4]. Los robots personales también pueden ser decisivos a la hora de apoyar el «capitalismo de vigilancia», una nueva forma de opresión social y económica basada en la vigilancia y en la manipulación de los usuarios, cuyos datos se capturan y venden[5]. Como ha defendido Carissa Véliz, los datos recogidos proporcionan a las empresas el poder para predecir e influir en nuestro comportamiento[6].

Usar robots personales como asistentes del hogar nos hace vulnerables a una forma de vigilancia de este tipo, junto con las amenazas que plantea para la privacidad y el poder. Para permitir una interacción a través de una interfaz de voz y el procesamiento del comportamiento humano, normalmente los robots personales están equipados con cámaras y micrófonos. Pueden grabar conversaciones y hacer vídeos. ¿Queremos que nuestras conversaciones privadas y nuestros momentos más íntimos sean capturados por un robot? ¿A dónde van a parar los datos? ¿Y qué se hace con ellos? En el caso de Jibo, los datos iban a la empresa del robot. Pero ¿quién tiene exactamente el acceso a los datos, y quién los ve y los escucha? ¿Qué ocurre posteriormente con los

datos? ¿Salen fuera de la empresa? Estas preocupaciones no tratan de escenarios de ciencia ficción. Ya tenemos muñecos y *smartphones* inteligentes que recogen datos y los envían a los servidores de la empresa para analizarlos. Si continúa esta tendencia, nos arriesgamos a avanzar hacia una sociedad de la vigilancia total en la que esta no ocurre solamente en la calle, sino también en la cocina y en el dormitorio. No solo los gobiernos (por ejemplo, a través de la policía, los servicios de inteligencia, etc.), sino también las empresas te pueden ver (o, al menos, siempre tienen la posibilidad de hacerlo) a través de robots, asistentes personales digitales y otros dispositivos. Las empresas de tecnología de la información capturan datos sobre ti, y pueden usar estos datos para predecir y manipular tu comportamiento, o venderlos a otras empresas que los quieran. ¿Qué pasa si los *hackers* obtienen acceso a los datos? Una vez que las tecnologías están conectadas a internet, no existe tal cosa como 100 % de seguridad.

> ¿Queremos que nuestras conversaciones privadas y nuestros más íntimos momentos sean capturados por un robot? ¿A dónde van a parar los datos?

Considérese el ejemplo de la Hello Barbie, una muñeca que graba y analiza lo que dicen los niños, y a continuación les responde[7]. Envía los datos a un centro de control de ToyTalk para procesarlos usando reconocimiento de voz. Esto permite que la muñeca pueda tener una «conversación» con los niños[8]. Los niños ganan un nuevo «amigo». Pero también se encuentran bajo vigilancia. Potencialmente, la muñeca se puede utilizar para hacer publicidad, basándose en el análisis de los datos del usuario. Además, como cualquier otro dispositivo que tiene conexión wifi y a internet, los *hackers* pueden obtener acceso a los datos personales de las conversaciones y, a través de Hello Barbie, a otros datos personales

almacenados en los dispositivos del hogar. Los robots personales equipados con reconocimiento de voz y capacidad de habla crearán problemas similares.

Los robots no solo recogen datos de audio: también pueden estar grabando tu cara y tu comportamiento. Combinados con la IA en la forma de reconocimiento facial y otros *softwares* inteligentes, los robots pueden crear nuevas oportunidades de vigilancia, tanto en la calle como en el lugar de trabajo, o en la intimidad del hogar. Si un robot puede reconocer tu cara, adivinar tus intenciones y, quizás, incluso leer tus emociones, cualquier humano que tenga acceso a esos datos y a los resultados de esos análisis de datos tendrá esta información sobre ti. Además, dada la interacción social, tales robots personales pueden proporcionar una información adicional sobre la personalidad del usuario del robot a través del comportamiento de dicho usuario. Como ha defendido Ryan Calo, el uso de un robot no es lo mismo que el uso de una lavadora. Cómo interactúa una persona con el robot, qué programa usa, etc., dice mucho sobre una persona: y todo esto está siendo grabado sin protecciones legales[9].

En segundo lugar, los robots móviles crean un problema adicional: la seguridad. Imagínese un robot que, por ejemplo, pueda causar daño a los niños cuando choca con ellos. Podría estar programado para detenerse cuando haya un humano cerca. Pero esto podría limitar su funcionamiento autónomo. Hironori Matsuzaki y Gesa Lindemann lo denominan la «paradoja de la autonomía-seguridad[10]». Como la introducción de robots industriales más inteligentes, los robots personales más autónomos nos invitan a reflexionar sobre nuevos conceptos de seguridad y sus aspectos éticos. Tampoco está claro dónde recae la responsabilidad cuando ocurre algo malo. También hay una cuestión ética relacionada con los humanos y los robots que comparten el mismo espacio: ¿deberían tener siempre prioridad los humanos (esto es, debería esperar siempre el robot por los humanos) o debería tener prioridad algunas veces? ¿Cómo deberían interactuar los robots con humanos (y viceversa) en situaciones particulares, como cuando ambos quieren pasar por la misma puerta al mismo tiempo? ¿Quién toma decisiones sobre estas cuestiones: los usuarios o la empresa del robot?

ENGAÑO, EL VALOR DE LAS RELACIONES HUMANAS REALES, DIFERENCIA Y SESGO DE GÉNERO Y RACIAL

La idea de usar robots asistentes para usuarios vulnerables como niños y ancianos ha suscitado un debate sobre si los robots pueden y deben usarse de

cuidadores (ver el capítulo siguiente), si los usuarios son engañados y si se respeta su dignidad. Esta preocupación se aplica no solo a los escenarios del tipo de los del *Mago de Oz*, en los que los usuarios son engañados de modo que piensen que el robot está operando de forma autónoma cuando no es así, sino también la idea de usar robots personales que realmente son autónomos e inteligentes.

Considérense los robots personales autónomos que podrían utilizar personas mayores para su cuidado o compañía. Robert y Linda Sparrow han argumentado que tales robots no pueden cumplir las necesidades sociales y emocionales de las personas mayores, existe un riesgo de menor contacto humano, y los robots no pueden proporcionar una interacción social genuina. No pueden proporcionar una verdadera amistad, amor o preocupación. Se trata de una simulación. Las personas pueden verse engañadas y pensar que el robot se preocupa de cuidar de ellas. Los Sparrow sostienen que este engaño es malo porque malinterpreta el mundo tal como es, y porque queremos dar amor y compañía reales, tener la creencia de que lo que se nos da no es suficiente: «lo que la mayoría de nosotros quiere de la vida es ser amados y que se preocupen por nosotros, tener amigos y compañeros, no *creer* meramente que somos amados o que los demás se preocupan por nosotros, o creer que tenemos amigos y compañeros, cuando, en realidad, esas creencias son falsas[11]». Incluso si nuestro bienestar subjetivo puede verse mejorado por el robot, lo que cuenta es el amor, la compañía, y el cuidado reales y objetivos. Robert Sparrow esboza un futuro distópico en el que los robots mantienen felices a las personas, pero no las respetan ni les dan muestras de agradecimiento. Él se imagina un edificio sin ventanas en el que los residentes ancianos están mirando pantallas y están siendo cuidados por robots, y en el que no hay humanos en sus proximidades[12]. Mientras que uno puede cuestionar filosóficamente la forma en que se plantea el problema, preguntarse cómo es realmente el cuidado robótico «virtual» e «ilusorio» (la interacción humano-robot es bastante real), y si el engaño es siempre y necesariamente malo en el cuidado, está claro que los robots no pueden ofrecer las mismas conexiones que los humanos, y que existe un peligro de manipulación y falta de respeto hacia las personas, así como a sus necesidades y dignidad humana[13].

Surgen preocupaciones similares en respuesta a la idea de usar robots como niñeras para niños pequeños. Noel y Amanda Sharkey plantean cuestiones sobre la privacidad y la seguridad (de nuevo, los robots pueden

grabar todo y pueden ser hackeados), pero también señalan otros problemas. Por ejemplo, ¿debería un robot tener permitido contener físicamente a un niño y, si es así, en qué circunstancias? Probablemente debería hacerlo si el niño quisiera cruzar una carretera peligrosa y se arriesgara a atravesar el tráfico. Sin embargo, ¿debería el robot evitar que el niño coja una galleta cuando le está prohibido por sus padres? ¿Dónde ponemos los límites? Si los robots asumen el trabajo «sucio» tal como cambiar pañales, los niños podrían perder la oportunidad de construir unas buenas relaciones con sus cuidadores (humanos). El cuidado robótico puede hacer peligrar el desarrollo mental y social del niño. Los autores señalan de nuevo el problema del engaño. Por supuesto, los niños juegan con muñecos y fingen durante el juego. Pero, al mismo tiempo, saben que los muñecos no están realmente vivos y que es un juego. Sin embargo, cuando un muñeco se sustituye por un robot autónomo, un niño puede creer que tiene una relación real con el robot y encariñarse con él[14].

Esta crítica se halla en la línea de la de Sherry Turkle, quien reconoce que los niños proyectan cuando interactúan con un muñeco (no robótico), pero mantiene que es diferente frente al caso en el que uno adquiere un compromiso con un robot personal, ya que en este caso el niño interactúa con el robot «como si estuviera frente a otra persona». Pero están siendo engañados: los robots no son personas. El robot simula. En lugar de interactuar con robots, argumenta Turkle, los niños necesitan «relaciones que les enseñan reciprocidad real, atención y empatía». Dicho de manera más general, los robots personales plantean la cuestión de si se pone en peligro la cualidad y el valor de las relaciones humanas. Los adultos también pueden desear compañía y esperarla de parte del robot. Sin embargo, afirma Turkle, los robots no la pueden proporcionar. Nos ofrecen un «como si»: nos hablan como si nos entendieran y se preocuparan por nosotros. Esto, sin embargo, no es una conversación auténtica. El problema no es solo que los robots nos engañen, sino también que, al final, nos hagan olvidarnos de lo que es una conversación auténtica. Ofrece el ejemplo del robot PARO, un robot con forma de una foca bebé que a veces se usaba en los asilos de ancianos: el robot parece que escucha y consuela a las personas mayores, pero los robots no pueden empatizar[15]. Esto es un engaño. Si interactuáramos principalmente con robots, ¿olvidaríamos en qué consiste tener una conversación real, y cómo son la compañía y amistad reales?

También se podría añadir una vaga perspectiva freudiana a la crítica de Turkle a los robots de compañía: nos gustan los robots en lugar de lo que realmente queremos (amor, amistad, buenas relaciones, etc.)[16]. El robot se convierte en un «fetiche»: un término que está en sí mismo relacionado con las palabras latinas «artificial» y «crear». Es un objeto fabricado por el hombre que (en este contexto) viene a encarnar nuestro deseo de valores humanos importantes. Lo imbuimos de la magia de lo que *realmente* buscamos y deseamos (en opinión de Sigmund Freud, deseo sexual). Sin embargo, de acuerdo con este análisis, esto es patológico. Mejor vayamos a por el objeto real.

Hablando sobre vida sexual, el caso de los robots sexuales ofrece otro ejemplo sobre cómo los robots personales pueden fomentar las críticas de que lo que los humanos son, dan y hacen con otros no puede ser reemplazado por robots ni por la interacción humano-robot. John Sullins duda de que los robots sexuales puedan satisfacer las necesidades físicas y emocionales de los seres humanos en lugar de meramente manipularlos. Sostiene que los robots no son capaces de crear una relación erótica en la que los amantes encuentran al otro en su complejidad, aprendan y se desarrollen. Los usuarios de robots sexuales tienen más bien el deseo de «borrar todo lo que en el amado no es un reflejo completo de las nociones preconcebidas que tiene el amante de lo que él o ella cree que quieren de un *partenaire*[17]». De nuevo, el argumento es que los robots no pueden proporcionar la cosa real. Quizás los robots sexuales sean una forma de evitar la complejidad y los desafíos de las relaciones humanas reales. Una vez más, parecen amenazar el desarrollo y el ejercicio de buenas relaciones humano-humano.

> La discusión sobre los robots sexuales es un buen ejemplo de cómo los robots personales, como robots sociales, nos hacen reflexionar sobre las relaciones *humanas*. Nos hacen pensar

sobre lo que son las relaciones humanas buenas.

No todo el mundo está de acuerdo con esta crítica. Algunos están más abiertos a las nuevas posibilidades ofrecidas por la tecnología. ¿Es demasiado convencional y conservadora la crítica que se ha mencionado? ¿Está el problema en el robot o en el usuario? Julie Carpenter ha sugerido que, a medida que la gente se acostumbrase a los robots y los acomodase culturalmente, los aceptaríamos y puede que se desarrollaran nuevas normas. Además, ha defendido que, en última instancia, las «recompensas, beneficios, cargas y repercusiones del apego emocional no descansan sobre el robot o su diseño, sino en el humano que interactúa con él[18]». Otros han defendido que los robots sexuales pueden ser una solución a la privación sexual o que son, básicamente, herramientas de masturbación (para una concepción general de los argumentos sobre los robots sexuales, véase John Danaher y Neil McArthur, *Robot Sex*). En cualquier caso, merece la pena señalar que siempre ha habido, no solo en la historia antigua, sino también en la moderna, relaciones complejas y, a veces, íntimas, entre los humanos y los objetos, como los muñecos eróticos, y algunos autores creen que no son necesariamente fetiches o sustitutos de otra cosa[19].

La discusión sobre los robots sexuales es un buen ejemplo de cómo los robots personales, como robots sociales, nos hacen reflexionar sobre las relaciones *humanas*. Nos hacen pensar sobre lo que son las relaciones humanas buenas y qué es el buen sexo. Incluso si se ponen en cuestión los conceptos mismos. ¿Qué son las relaciones humanas (buenas)? ¿Qué es el sexo? ¿Qué *es* «la cosa real»?

Por ejemplo, la discusión de Carpenter sobre los robots sexuales abarca las nociones de sexualidad, apego, amor, relaciones románticas y teoría sobre las relaciones humano-humano. El caso de los robots sexuales también nos hace reflexionar sobre las relaciones entre hombres y mujeres y, de forma más general, sobre cuestiones de género. Kathleen Richardson, que lanzó una campaña contra los robots sexuales, ha mantenido que su uso proyecta una imagen de trato con el otro en la que el otro (a menudo una mujer) se «convierte en una cosa». De esta forma, los robots sexuales conllevan el

riesgo de que se refuercen relaciones de poder asimétricas en las que la subjetividad humana de uno de los miembros de la pareja no se vea reconocida: la empatía se «apaga[20]». Se podría concluir que los robots no solo tienen un impacto directo en la ética como en las violaciones de la privacidad y la seguridad: a un nivel más profundo, están conectados con la forma en que pensamos y hacemos las cosas con los demás.

Contra estas objeciones se podría proponer que, desafortunadamente, los humanos no han esperado a que los robots personales proporcionen unos malos cuidados, fracasen a la hora de empatizar con otros seres humanos o traten a los demás como cosas. No se debería idealizar lo que los humanos hacen a los humanos. Aun así, el hecho de que los humanos no siempre estén a la altura de estándares éticos elevados no es una contraobjeción válida, ya que no socava los argumentos contra (estos usos de) los robots personales como tales. Como mucho, puede fomentar mejores relaciones humano-humano en general, con y sin robots. La acusación de engaño, sin embargo, se merece una mayor discusión.

Los argumentos contra los robots personales hacen a menudo que parezca que el engaño es necesariamente malo. Pero ¿es así? Mucho de ello depende de cómo definamos el engaño. Si engaño significa discurso engañoso y esto, a su vez, se define como mentira maliciosa, entonces la mayoría de la gente estará de acuerdo en que el engaño es malo. Sin embargo, si el engaño en el discurso se define en términos de lo que Alistar Isaac y Will Bridewell describen como «una característica habitual de la interacción humana cotidiana», que abarca «mentirijillas» y despistes, entonces las cosas se ven de otra manera. Ellos afirman que este tipo de comunicación cumple funciones socialmente benéficas y los robots genuinamente sociales «necesitarán la capacidad para participar en el mercado humano del engaño[21]». En otras palabras, el razonamiento es que, si el engaño es parte de lo que significa funcionar como una entidad social, y si los robots deben convertirse en tales entidades, entonces necesitan ser capaces de engañar. Sin embargo, esto deja sin respuesta cuándo (en qué situaciones) las mentirijillas son moralmente aceptables. Además, la premisa de que debemos construir entidades sociales artificiales también necesita más apoyo. Por ejemplo, ¿para qué necesitamos robots que nos hablen?

En segundo lugar, el engaño puede ser definido como la creación de algo ilusorio. ¿La creación de fantasias, que podría decirse es lo que hacen los desarrolladores de estos robots, es necesariamente mala? Por ejemplo, si se

crea una fantasía en la que el robot es un ser vivo o un amigo, ¿es esto algo malo? Seguramente no tengamos un problema con lo ilusorio cuando vamos al teatro o a un espectáculo de magia. Estamos contentos de ser «engañados» durante el espectáculo. Antes y después, sin embargo (y, quizás, durante), al mismo tiempo sabemos que lo que estamos viendo no es real. Ocurre lo que se conoce como suspensión de la incredulidad: la gente acepta temporalmente un espectáculo o historia como realidad para estar entretenida. Esto, aplicado a los robots personales, significaría que la gente suspendería la incredulidad de forma temporal durante la interacción con el robot para que, así, el robot pueda hacer su «magia». Éticamente hablando se podría exigir entonces a los diseñadores y a aquellos que ofrecen robots a los usuarios, que los usuarios sean conscientes de que el robot está creando un mundo ilusorio, que lo que ocurre en la interacción humano-robot es «como si» y fingido. Este es un desafío para los diseñadores y los desarrolladores, pero también para los padres, los trabajadores sanitarios y otros que ofrecen el robot a aquellos que se suponen que deben cuidar. Se necesita un tipo de honestidad sobre lo que es realmente el robot y lo que puede proporcionar. Esto iría contra gran parte de la publicidad actual de los robots personales, que a menudo se venden como tu «amigo», que te proporciona «compañía» y que te facilita mantener «conversaciones», y así sucesivamente.

Por otro lado, la línea que separa la realidad y la ilusión no está tan clara como pudiera parecer a primera vista. Como se ha dicho, la interacción con el robot es real, y las emociones por parte del humano son reales durante tales «actuaciones» humano-robot. La interacción humano-robot no tiene lugar en un mundo de fantasía. Considerada desde esta perspectiva, una distinción platónica estricta entre la ilusión y la realidad parece difícil de mantener. En lugar de existir dos mundos (uno real y otro falso), las cosas son más complejas. Podríamos decir que hay una actuación por parte del robot y sus usuarios, y que juntos crean algo que es ilusión y realidad al mismo tiempo, de forma similar a lo que ocurre en el escenario de un teatro o de un espectáculo de magia. Además, tomar en serio el aspecto interactivo implica reconocer que no solo el robot o sus creadores crean algo ilusorio: los usuarios al menos *co-crean* la ilusión en la medida en que participan en la interacción y la actuación robótica como espectadores y como co-actores potenciales. Como sugiere el término «suspensión de la incredulidad», normalmente los usuarios *quieren* ser engañados y engañarse ellos mismos. Quieren magia. Parece, por lo tanto, demasiado simplista representar a los

diseñadores y a los cuidadores como necesaria y completamente culpables, e interpretar al usuario como alguien completamente inocente. Al menos parece haber una diferencia crucial entre, por ejemplo, los adultos y los niños pequeños. La mayoría de los adultos pueden ser co-creadores mucho menos inocentes de la ilusión cuando usan los robots personales (aceptan la ilusión temporalmente, pero saben que es una ilusión), mientras que es razonable tener una preocupación ética adicional en el caso de los usuarios vulnerables como los niños pequeños o, digamos, personas con discapacidades cognitivas que pueden no saber en absoluto qué es estar ante algo ilusorio. La ética de la creación de ilusiones y, por lo tanto, la ética de los robots personales es más compleja que lo aquí presentado[22].

Más allá de cuestiones relacionadas con el engaño y las relaciones íntimas humanas, es cuestionable si el uso de los robots sociales personales respeta la diferencia de forma suficiente: diferencias culturales y personales. Si los robots no son simplemente artefactos, sino que también están ligados a sus entornos sociales y culturales, la consecuencia es que esto conlleva implicaciones éticas. Los robots se *desarrollan* en una sociedad particular y en una parte del mundo, como los Estados Unidos o Japón, y la forma en que se presentan y se comportan es probable que esté influenciada por los valores y la forma de vida de esa sociedad y cultura. Esto puede estar en tensión con la sociedad, la cultura y los valores del usuario y del entorno en el que se *usa* el robot. Por ejemplo, un robot y una IA diseñados para operar en una cultura individualista (que haya bebido, por ejemplo, de los textos y conversaciones de internet occidentales), puede que no funcione siempre bien en un entorno cultural más comunal o colectivista. Además, el diseño y el desarrollo de los robots puede que den por supuesto un tipo de usuario estándar universal, pero todos somos personas diferentes, con necesidades y expectativas potencialmente diferentes, diferentes cuerpos y géneros, y diferentes historias personales. Considérese, por ejemplo, los robots personales cuya interacción con los humanos depende casi exclusivamente de la interfaz de voz, y que no pueden ser usados por aquellos que tienen una discapacidad auditiva. O los robots que suponen una distribución de género binaria en sus comunicaciones, hablando como si no existieran las personas transgénero u otros que no encajen en estas categorías. También existen diferentes expectativas sobre lo que es el buen cuidado: algunas personas quieren el mínimo contacto social y estar solas en algunas situaciones, otros quieren hablar mucho y esperan empatía, y así sucesivamente. ¿Cómo podrán lidiar con esto los robots

personales? Se podría usar la IA para hacer una adaptación de la inteligencia del robot al usuario y a su cultura. Pero es probable que se mantenga la fricción potencial. Los humanos pueden aprender a lidiar con la diferencia y respetar a otros que son diferentes de ellos mismos. ¿Pueden las máquinas ofrecer ese tipo de respeto a los humanos, dado que no pueden preocuparse de nada en absoluto? Quizás puedan *simular* el respeto por la diferencia. ¿Sería eso suficiente? De nuevo, esta cuestión hace surgir objeciones con respecto al engaño junto con las violaciones de las necesidades y dignidad humanas.

Finalmente, una cuestión relacionada con la sociedad y la cultura es que los robots personales pueden tener comportamientos, significados e implicaciones más amplias que creen, mantengan o incrementen los sesgos, como los de género o los raciales y, de forma más general, causen impacto en la justicia y la igualdad de una sociedad específica. Ya he mencionado que los robots sexuales plantean cuestiones de género. Pero otros robots también pueden ser problemáticos en términos de género. Los artefactos tecnológicos como los robots pueden ser sexistas, en el sentido de si están conectados con significados de género normativamente relevantes que estén disponibles en una cultura o sociedad particular, o con lo que he llamado «gramáticas de género» y «juegos de género[23]». Por ejemplo, un robot de limpieza o del hogar con la forma de una mujer es problemático, ya que perpetúa las percepciones y prácticas que categorizan a las mujeres como pertenecientes a esos roles. De forma más general, los robots están relacionados con nuestras expectativas culturales en sus roles como sirvientes, amigos, esclavos y así sucesivamente, y estos roles están «cargados con los estereotipos del usuario», como señalan Carpenter y sus colegas[24]. Incluso si usualmente no se pretende por parte de los diseñadores, los robots tienen este efecto, ya que siempre existe un vínculo con la sociedad. En este sentido, los robots no son nunca meras cosas.

El género de los robots es, así pues, problemático, ya que puede «perpetuar aspectos de ciertos roles humano-humano y de las ideologías que los acompañan[25]». Esto también es cierto para la discriminación y el racismo. Como ha argumentado Ruha Benjamin, las tecnologías de la automatización, como los robots y la IA pueden profundizar de forma involuntaria la discriminación y tener efectos racistas[26]. En el contexto de los EE. UU., por ejemplo, esto puede significar que con tales tecnologías se corra el riesgo de reforzar una historia de opresión y promocionar la supremacía blanca. A pesar de no tratarse de robots como tales, el aprendizaje

automático usado en los concursos de belleza con jueces de IA puede conducir al sesgo racial. Otros ejemplos famosos incluyen *software* de etiquetado fotográfico [*photo-tagging*] que clasificaba a los usuarios de piel oscura como gorilas, o los algoritmos de predicción del crimen que estaban sesgados en contra de personas que viven en los barrios negros. Pero Benjamin también señala otras formas en las que los robots pueden ser problemáticos en términos de la raza y preocupaciones sociales relacionadas: la historia de la robótica con su deseo de tener esclavos personales, el uso de los robots por parte de la policía para la vigilancia y la fuerza letal contra poblaciones raciales concretas (un tema extremadamente relevante cuando estoy escribiendo este libro), sensores que solo reconocen un pequeño espectro de tonos de piel, y desigualdades en la fuerza de trabajo tecnológico en cuanto a cuestiones raciales y de género. Considérense, por ejemplo, los trabajadores manuales mal pagados en el Sur global, a menudo mujeres, que trabajan para creadores y empresarios blancos posicionados en la cúspide de la jerarquía. A pesar de que normalmente no exista una intención racista de parte de los diseñadores y desarrolladores de los robots, existe un riesgo significativo de que los robots perpetúen una historia de dominación y racismo, continuando con lo que llamaría, por analogía con los «juegos de género», «juegos raciales» concretos[27]. Una sociedad particular ya tiene formas específicas de lidiar con la raza, y esos contextos sociales y culturales más amplios, incluyendo sus (normalmente no escritas) reglas y problemas que hacen surgir, están reflejados, ejecutados y mantenidos en el uso de tecnologías como los robots y la IA. En este sentido, los robots pueden ser racistas, como señala Benjamin, o como diría yo, las actuaciones y los juegos en los que los robots desempeñan un papel pueden ser racistas, y los robots pueden continuar y promover estos juegos.

Según queda claro en los ejemplos, este análisis en términos de juegos de género y raza no es aplicable solamente a los robots personal/es, sino también a los robots en general. Pero dada/s su/s función/es social/es, los robots personales, sobre todo cuando tienen una apariencia humana, necesitan atención ética especial cuando se trata de cuestiones como el género y la raza. El problema con los robots con apariencia humana no es solo que puedan parecer humanos (y, por tanto, plantear el problema del engaño). El problema es que tienen apariencia de (categorías de) humanos *particulares*, y pueden tener efectos potencialmente dañinos sobre (categorías de) humanos

particulares en la medida en que interactúan con humanos (particulares) y tienen efectos sociales más amplios.

Para finalizar, los robots personales (en tanto que robots sociales) plantean cuestiones metafísicas, éticas, sociales y existenciales que resultan ser cruciales. La discusión nos conduce fácilmente más allá de la «ética» si se define ese campo de forma estrecha, como si tratara de lo que es lo correcto y lo incorrecto. También nos hace pensar sobre lo que es real, lo que valoramos y lo que queremos, los problemas de la sociedad y lo que significa ser humano. Como observa Carpenter, experimentando y reflexionando sobre los robots podemos obtener «perspectivas sobre lo que significa ser humano o no humano», y esto «puede hacernos reflexionar sobre nosotros mismos y lo que realmente queremos de los demás[28]».

LA ÉTICA DE LOS ROBOTS PERSONALES MÁS ALLÁ DE LOS HUMANOS

La discusión previa aceptaba que en la ética de los robots solo debemos preocuparnos por los humanos. Pero hay más cuestiones éticas adicionales que están relacionadas con el hecho de que los robots personales operan en entornos en los que no solo hay humanos, sino también *animales*. ¿Cuáles son las implicaciones éticas de tales robots, por ejemplo, para las mascotas? ¿Cómo reaccionan, por ejemplo, los gatos y los perros ante los robots y viceversa, y en qué consiste una forma ética con la que los robots se relacionen con ellos? Hay, digamos, problemas de seguridad. Se podría exigir que los robots que se mueven por la casa no deberían hacer daño a los animales. Sin embargo, dado el carácter social y conversacional de los robots, existe una cuestión más amplia: ¿cómo intervienen, deberían intervenir e influir los robots en el tejido social de las relaciones entre humanos y animales?

¿Cuáles serían y deberían ser su lugar y papel? ¿Cambiarían el lugar y el papel de las mascotas cuando se introduzcan los robots en el hogar? ¿Cómo serán las interacciones robot-animal? ¿Cómo debemos dar forma a estas interaccionas sociales y redes mixtas humano-no humano? Aceptando que la ética deba tomar en consideración no solo los intereses, necesidades y el valor de los humanos, sino también de los animales, estos son desafíos pendientes de los robots personales, desafíos que raramente se abordan.

Además, más allá del impacto sobre nuestras relaciones con los animales, está la cuestión del impacto *medioambiental* de los robots. No solo los robots industriales, sino también los robots personales tendrán un impacto en el

entorno natural por la forma en que se producen, usan y se gestionan tras su uso. Por ejemplo, la ética de los robots personales debería preguntarse, cuando se rompe el robot (o se apaga por la empresa), ¿qué ocurre con el *hardware*? ¿Acaba como residuo? ¿Se reciclará? ¿Cómo debería producirse? ¿Cuánta energía consume? ¿Qué materiales deberían usarse? ¿Cuál es la huella de carbono? ¿Y cuáles son las responsabilidades con respecto a estas cuestiones de parte de los diseñadores, compañías, usuarios y autoridades públicas del robot? El caso de los *smartphones,* los dispositivos personales que ya tenemos la mayoría de nosotros, no es prometedor: tienen ciclos de vida cortos y raramente se reciclan, y su producción, operación e infraestructuras implican un gran consumo de energía (por ejemplo, en los servidores en funcionamiento), emisiones de CO_2, y extracción de minerales raros. ¿Ocurrirá lo mismo con nuestros «amigos» los robots cuando la empresa que los fabrique o venda nos aliente a comprar uno nuevo cada dos años? ¿Y cuál será la huella de carbono de la infraestructura necesaria para la energía y la IA que impulsa a todos los robots?

En el capítulo final volveré sobre estas cuestiones acerca de una ética de los robots que va más allá de lo humano.

La discusión nos conduce fácilmente más allá de la «ética»: también nos hace pensar sobre lo que es real, lo que valoramos, los problemas de la sociedad y lo que significa ser humano.

Capítulo 4

Robots asistenciales, pericia y la calidad de la atención sanitaria

En la película de ciencia ficción de 2012 *Un amigo para Frank*, un hombre mayor con demencia recibe un robot de su hijo, que está cansado de hacer visitas regulares a su padre. El robot asume las tareas domésticas y proporciona cuidados terapéuticos, monitorizando las rutinas y comidas diarias de Frank, pero animándolo también a que practique la jardinería para mantener sus habilidades cognitivas. A pesar de que a Frank no le gusta el robot al principio, empieza enseguida a valorarlo y lo usa para sus propios fines. El robot se convierte en un amigo, y deja de ser un sirviente o un esclavo. Su hija, sin embargo, cree que la máquina es éticamente cuestionable. Hay incluso un movimiento social contra ella. ¿Son los robots asistenciales una forma de abandonar a los mayores? Por ejemplo, si son posibles el control y la monitorización remotos, ¿es probable que las personas visiten cada vez menos a sus parientes ancianos? ¿Y está bien depender completamente de la tecnología? ¿Puede el robot ayudar en las tareas domésticas sin que las personas se vuelvan totalmente dependientes? La hija de Frank dice cuándo está de acuerdo en que se use el robot. Pero si vemos lo que ha pasado con los teléfonos móviles, puede que la gente se vuelva *muy* dependiente de la tecnología.

Este escenario no es muy lejano. El Gobierno japonés cree que los robots son una forma de lidiar con la creciente población anciana[1]. Un razonamiento relacionado a favor de usar los robots asistenciales es ahorrar en costes de personal. Ya se están probando y usando robots en la atención sanitaria, no solo en Japón, sino en otras partes del mundo. Por ejemplo, el robot PARO, creado por la empresa japonesa AIST, se comercializa como un «robot terapéutico» que puede usarse, y ya se ha usado, en hospitales y centros de atención para relajar a los pacientes, así como para estimular la

interacción social entre pacientes, y entre pacientes y cuidadores. Parece una foca bebé y responde a las caricias y a la voz. «Al interaccionar con la gente, PARO responde como si estuviera vivo, moviendo su cabeza y patas, emitiendo sonidos y mostrando el comportamiento que uno prefiera. PARO también imita los sonidos de una foca arpa bebé real»[2]. Puede ser cierto que con ello se estimule la interacción social. Pero ¿puede haber algo malo al usar esos robots de esta forma? ¿Es engañoso, como nos preguntábamos en el capítulo anterior? ¿Es humillante? ¿Se está con ello infantilizando a la gente? ¿O deberían subrayarse los beneficios terapéuticos? ¿Cuáles son las cuestiones éticas que se plantean al usar robots en la atención sanitaria?

LOS ROBOTS EN LA ATENCIÓN SANITARIA

El robot PAROT ilustra cómo se pueden usar los robots en el cuidado de los ancianos. Los robots pueden usarse también para cuidados de enfermería en general. Por ejemplo, en un entorno de hospital, se pueden usar para transportar suministros médicos y comida, entregar medicinas o mover pacientes. Ya hay robots que ayudan a las enfermeras a realizar tareas que no tienen que ver con los cuidados propiamente dichos, como llevar suministros y medicinas, tal como hace el robot Moxi en los Estados Unidos[3]. O pueden desempeñar un papel en la llamada telemedicina: los pacientes que viven en áreas remotas, las personas con necesidad urgente de ver a un doctor, o que reciben atención sanitaria en casa, pueden tener un diagnóstico a través de un robot que está controlado de forma remota. Los robots también pueden (ayudar a) monitorizar personas que están en casa y quizás intervenir si algo no va bien. Para las personas mayores, pongamos por caso, un robot podría comprobar sus constantes vitales después de una caída. Además de la asistencia y la monitorización, se afirma que algunos robots pueden proporcionar compañía (ver la discusión en el capítulo anterior).

Pero los robots también se pueden usar en otras áreas de la atención sanitaria. En la cirugía asistida por robots, el cirujano la lleva a cabo de forma remota a través de control informático o usando un telemanipulador. En este caso, el cirujano no necesita un contacto directo con el cuerpo del paciente: el robot realiza la cirugía real. La tecnología se utiliza principalmente para la denominada cirugía mínimamente invasiva, que significa que solo se necesita una pequeña incisión. Por ejemplo, en el caso del robot cirujano Da Vinci, este tiene montados en el brazo robótico ciertos instrumentos y una cámara 3D que son controlados por el cirujano, el cual controla el brazo (todo a la

vez) mediante una consola con un monitor que ofrece una buena vista de lo que ocurre dentro del paciente. Para el cirujano, hacer esto requiere un entrenamiento específico y aprender nuevas habilidades, pero una vez aprendido, el sistema facilita un control incrementado y, en consecuencia, cirugía más precisa. Para el paciente, una cirugía mínimamente invasiva y más precisa implica que esta sea menos traumática y haya una recuperación más rápida[4]. Hoy en día el uso de estos robots está ampliamente aceptado por los pacientes, a pesar de que existen algunos problemas éticos. Por ejemplo, los pacientes necesitan estar suficientemente informados, el coste de la tecnología es relativamente alto, y cuando algo sale mal, no está claro cómo distribuir la responsabilidad legal y moral, ya que hay muchas partes implicadas, incluyendo, al menos, al cirujano, al hospital y al fabricante (de forma similar a lo que sucede con la responsabilidad en el caso de los coches autónomos, ver el capítulo siguiente)[5].

El ejemplo de los robots cirujanos muestra que los robots en la asistencia sanitaria no toman necesariamente la forma de robots humanoides que operan de forma independiente, esto es, no tienen la imagen que nos llega de las películas de ciencia ficción. Por supuesto, *hay* robots que se parecen los humanos o a los animales. Consideremos de nuevo a PARO. Pero hay, por ejemplo, robots exoesqueletos. El robot HAL se usa en rehabilitación y ayuda a las personas a volver a caminar[6]. Además, los robots son parte de sistemas sociotécnicos mayores. No existe solamente una «interacción diádica entre un humano y un robot»[7]. Los robots impactan en la atención sanitaria en su conjunto, cambiando, por ejemplo, el papel del personal de atención sanitaria, y con implicaciones para la educación y la pericia.

Más controvertido que los robots cirujanos es el uso de los robots de asistencia social en la atención sanitaria mental, que se han utilizado, por ejemplo, en el cuidado de demencia y en la terapia para niños autistas. En el último caso, el robot puede usarse para reemplazar al terapeuta, desempeñar un papel de mediación, o asistir al terapeuta. El primer papel se ha probado, por ejemplo, en el proyecto europeo DREAM, que usa un robot con autonomía supervisada. Cuando no está bajo control remoto (Mago de Oz), el robot opera de forma autónoma durante algún tiempo, pero el terapeuta está aún presente[8]. Esto plantea todo tipo de cuestiones con respecto a la apariencia y al comportamiento del robot (por ejemplo, ¿es aceptable que el robot, como el robot NAO, tenga la apariencia de un humano?), la

responsabilidad, la confianza y la privacidad[9]. Como en el caso de los robots asistenciales, la preocupación es que los robots se usen simplemente para ahorrar dinero sin un beneficio claro para el paciente. En este caso, estas cuestiones fueron investigadas y discutidas como parte del proyecto de investigación[10]. Pero a menudo falta este componente ético en el desarrollo del robot.

PROBLEMAS ÉTICOS

En general, los robots en la asistencia sanitaria plantean diferentes problemas éticos. Presentaré un resumen haciendo hincapié sobre los robots inteligentes y autónomos en tareas de asistencia en el contexto de la enfermería y la atención de ancianos.

En primer lugar, como todas las tecnologías en red, existen de nuevo problemas ligados a la privacidad, a la protección de datos y a la vigilancia. ¿Qué datos recoge el robot, cómo se almacenan, quién tiene acceso a ellos, quién los posee, y qué ocurre con los datos ahora y en el futuro? Tanto si los robots se usan en casa (ver también el capítulo previo) o en centros de atención y hogares para ancianos, la tecnología permite una vigilancia continua. ¿Es esto éticamente aceptable? ¿Qué ocurre si la gente aún no puede o ya no puede otorgar su consentimiento?

En segundo lugar, como sucede con los robots industriales, está también la cuestión ligada a las implicaciones para la fuerza de trabajo. Si el robot asume las tareas de los humanos (y hasta qué punto), serán necesarios menos humanos: el resultado es el desempleo. Puede considerarse que (categorías de) humanos específicas que trabajan en puestos de atención sanitaria ya no son necesarios y, por lo tanto, verse especialmente afectados, incluso de forma involuntaria, como las mujeres y los inmigrantes. Como se ha indicado, una de las principales razones ofrecidas para la introducción de los robots es ahorrar costes de personal. Pero las soluciones tecnológicas no son la única o la necesariamente mejor forma de abordar los desafíos demográficos. Se podrían emplear más trabajadores de atención sanitaria. Existe la sospecha de que estas concepciones de la robótica en la atención sanitaria tienen menos que ver con mejorar la calidad del servicio, y más con justificar las inversiones públicas y privadas en el sector tecnológico, y la sustitución de empleados humanos en este sector para así reducir costes de trabajo (y, en algunos casos, no emplear a inmigrantes; se ha sugerido con frecuencia que esto desempeña un papel importante en el contexto japonés). Por

consiguiente, estos no son simples problemas económicos o laborales, sino que están también relacionados con cuestiones relativas a la justicia y a la igualdad.

En tercer lugar, junto a estos problemas «macro» relacionados con la sustitución y su impacto socioeconómico dada la ya mencionada integración sistémica de los robots, existen más problemas «micro» y «meso» que surgen de la autonomía del robot. ¿Cuál es el impacto sobre las organizaciones, las prácticas y las interacciones? La sustitución es, de nuevo, una preocupación clave en este caso. ¿Cuál será exactamente el papel de los robots? ¿Asumirán trabajos al completo o solamente tareas específicas? ¿Implican estas tareas colaboración con humanos? ¿Cuál es el papel resultante de las enfermeras, los doctores y los cirujanos? ¿Qué implica esto para las organizaciones? ¿Cuál es el impacto sobre los compañeros, parientes y amigos? Existe la preocupación de que el uso de robots de asistencia conduzca a menos contacto humano[11].

Además, ¿quién es responsable cuando algo sale mal? Existe una «brecha de responsabilidad» en el sentido de que los robots se vuelven cada vez más autónomos y se les asignan más tareas, pero los robots carecen de la capacidad de agencia moral[12]. ¿Cómo seguirán siendo responsables los humanos si no tienen control suficiente? ¿Cómo se puede asegurar la atribución de responsabilidad (a los humanos) si el robot no está supervisado? (Ver también el capítulo siguiente). Este problema se puede resolver teniendo robots supervisados. Pero, entonces, ¿por qué usar robots (si el objetivo es reducir el coste de personal), y hasta qué punto es responsable el supervisor cuando el robot hace algo mal, dado que este ha sido desarrollado por otras personas?

Más allá de la cuestión de los costes, una razón por la que podríamos necesitar a los robots es para hacer que ciertas tareas sean más sencillas para el personal sanitario. Los robots no se usarían entonces para sustituir, sino para ayudar y mejorar. Considérese a una enfermera que tiene que levantar de la cama a un paciente: un robot puede ser útil en este caso, ya que evita problemas de espalda a la enfermera. Considérese de nuevo el caso del robot cirujano o, incluso, el de atención mental supervisada. En tales casos la asistencia y la mejora parecen ser mucho menos problemáticas que, digamos, las perspectivas de un robot que distribuye medicinas a los pacientes o que asume el trabajo integral de enfermería. La discusión ética depende, así, mucho de las tareas y papeles que tengan los robots.

En cuarto lugar, hay problemas que conciernen a la interacción humana que ya se han mencionado en el capítulo anterior y en la introducción, como el engaño, la confianza y la libertad. Si el robot finge ser un animal o un humano, ¿es este un caso de engaño, y está mal *per se*? Considérese de nuevo el caso de PARO y, de forma más general, la discusión iniciada por los Sparrow y los Sharkey[13]. ¿La gente que está siendo cuidada por robots personales está viviendo una ilusión? ¿Están siendo engañados? ¿Pueden los pacientes y el personal sanitario confiar en los robots? ¿Es «confianza» la palabra correcta? ¿O se usa simplemente como un término que suena mejor que «aceptación», que es de lo que tienden a preocuparse los investigadores y la industria robótica? ¿Se debería permitir a los robots restringir la libertad de, digamos, una persona anciana con demencia o de una persona con tendencias suicidas?

Los robots en contextos de atención también pueden generar problemas de género. Ya he introducido este problema en el capítulo anterior y mencionado el aspecto potencial de género del riesgo de desempleo, pero también hay cuestiones de género en el nexo donde se encuentran lo micro y lo macro: esto es, en la interacción humano-robot en cuanto que está ligada a las expectativas y normas culturales. Por ejemplo, incluso hoy la enfermería se percibe normalmente como una ocupación para mujeres. Los robots asistenciales con voz y/o apariencia de mujer pueden reforzar este estereotipo. La discusión sobre los robots en la atención revela, así, supuestos (incluyendo estereotipos) sobre mujeres, enfermeras y atención (sanitaria).

> Las soluciones tecnológicas no son la única ni necesariamente la mejor forma de abordar los desafíos demográficos.

En quinto lugar, relacionado con la preocupación de que hay menos contacto humano se halla el argumento de que los robots no pueden

proporcionar un tipo de cuidado «cálido», esto es, no pueden proporcionar un tipo de cuidado que responda a las necesidades sociales y emocionales de los pacientes. La idea es que los pacientes, los ancianos, etc., no solo necesitan cuidado técnico «frío» en el sentido de procedimientos médicos o medicinas, sino que también quieren tener una conversación, que se reconozcan sus necesidades emocionales, que se respete su dignidad, etc., necesitan mantener conversaciones significativas y apoyo emocional. Los robots no pueden proporcionar este tipo de cuidado, ya que carecen de emociones y de consciencia. Por lo tanto, se concluye, los robots no se deberían usar en la atención sanitaria. Contra este argumento se ha señalado que la atención sanitaria actual es más bien fría e, incluso, que debería ser así, ya que, en primer lugar, demasiada implicación emocional haría imposible el trabajo del personal sanitario y, en segundo lugar, no todos los pacientes quieren un cuidado cálido: algunos quieren un tipo de cuidado más distante. Por ejemplo, se dice que algunos prefieren una máquina que los lave en lugar de un ser humano porque es menos vergonzoso y más dignificante. Un defensor del cuidado «cálido» podría responder que el hecho de que el cuidado actual sea frío no es un argumento válido contra el cuidado frío como tal. Está claro que también hay problemas con la atención sanitaria actual, pero esto no implica por sí solo que el uso de robots sea aceptable. Además, a pesar de que es cierto que demasiada implicación puede ser problemática para el personal sanitario, esto no dice nada en contra de incluir *algún* grado de calidez: esto es, alguna respuesta a las necesidades emocionales y alguna dimensión social. Mientras que algunas personas piensan que prefieren un cuidado frío, es importante atender a la gente que quiere un tipo de cuidado diferente. ¿Qué pasa si la gente que prefiere un cuidado frío está en una situación en la que ya no tienen familia (o que su familia ya no les visita de forma regular) y el personal sanitario son los únicos seres humanos que conocen?

En sexto lugar, cuando los robots son usados por adultos, como, por ejemplo, los ancianos, existe la preocupación de que este uso pueda no respetar la dignidad de las personas y que, en cambio, las infantilice. Considérese de nuevo el robot PARO. Ofrecer este robot a los ancianos con demencia puede llevarse a cabo de una forma que los trate como a niños.

O, peor aún, puede que sean tratados como cosas en lugar de como personas. Uno de los problemas relacionados con el tratamiento de las personas con demencia es lo que Tom Kitwood ha denominado «objetificación»: «tratar a una persona como si fuera un montón de materia

muerta[14]». Dicho de forma más general, es importante que los robots no amenacen la dignidad de las personas; y por supuesto, la dignidad humana se respeta *en general* en la atención sanitaria, con o sin robots. Los Sharkey ofrecen una historia escalofriante sobre una persona anciana que no consigue una bacinilla para orinar porque, como le cuenta la enfermera a su mujer, las sábanas se cambiarán de todas formas. El problema relacionado con la gente con demencia también plantea la cuestión del tipo de imagen que tenemos de aquellos que necesitan cuidados, y cómo lidiar con usuarios vulnerables mientras tenemos en consideración sus condiciones y necesidades diversas (junto con las necesidades del personal sanitario).

Estas cuestiones que tienen que ver con el engaño, la confianza, el cuidado cálido y la dignidad humana están relacionadas con otras más generales que deberíamos plantearnos con respecto a la introducción de los robots en la asistencia sanitaria. ¿Qué es un buen cuidado? ¿Lo pueden proporcionar los robots?

¿QUÉ ES UN BUEN CUIDADO?

El argumento de los Sparrow contra el uso de robots de asistencia social en la atención sanitaria no trata solo sobre engaño, sino, de forma más general, de la calidad del cuidado. Se podría responder que lo que realmente importa no es tanto (la ausencia de) engaño, confianza o cuidado cálido como tales, sino un *buen* cuidado. En particular, contra la noción de que es aceptable engañar a la gente si eso los hace sentirse mejor, los Sparrow mantienen que lo que cuenta no es tanto si los pacientes se *sienten* bien, subjetivamente, sino también si se cumplen los criterios objetivos de un buen cuidado[15]. Pero ¿qué es un buen cuidado?

Comencemos con la afirmación de que el sentimiento y la experiencia del cuidado no es lo *único* que cuenta. Un dispositivo que se puede utilizar para argumentar a favor de este punto de vista (y, por tanto, apoyar aún más lo que ya señalaron los Sparrow) es lo que he llamado la «máquina de experiencia de cuidado», modelada sobre el experimento mental de la «máquina de experiencias» de Robert Nozick[16]. Nozick escribe:

> Supongamos que existe una máquina de experiencias que nos ofreciera cualquier experiencia que deseáramos. Neuropsicólogos superdotados podrían estimular tu cerebro de forma que pensaras y sintieras que estás escribiendo una gran novela, o haciendo una amistad, o leyendo un libro interesante. Estarías todo el tiempo flotando en un contenedor, con

> electrodos conectados a tu cerebro. ¿Deberías conectarte a esta máquina de por vida, preprogramando tus deseos vitales? […] ¿Te conectarías?

¿Qué más nos puede importar a nosotros, que cómo se sienten nuestras vidas desde dentro?[17].

Supongamos entonces que, si los robots pueden ofrecer a los pacientes la experiencia de recibir un buen cuidado, ¿tal cuidado «virtual», usando el término de los Sparrow, sería éticamente aceptable? El experimento mental está diseñado, por supuesto, para hacernos responder «No, el buen cuidado subjetivo no es suficiente. Necesitamos un cuidado real y un buen cuidado». El asunto no es, ni debería ser, que la experiencia del cuidado no es relevante en absoluto. Se podría aun afirmar que la experiencia cuenta, pero que se deben cumplir los criterios objetivos del buen cuidado al mismo tiempo. Los criterios subjetivos que son a la vez criterios objetivos constituyen las condiciones necesarias, pero ninguna condición es por sí misma suficiente para ser un buen cuidado. Entonces, ¿qué es *objetivamente* un buen cuidado?

> Cuando los robots son usados por adultos, como, por ejemplo, los ancianos, existe la preocupación de que este uso no pueda respetar la dignidad de las personas y qué, en cambio, las infantilice.

Los criterios objetivos para valorar el buen cuidado pueden derivarse de diferentes fuentes. Los Sharkey proponen evaluar la cuestión relacionada con los robots asistenciales usando los derechos humanos y los valores humanos compartidos[18]. Considérense los derechos a la salud y al bienestar, la privacidad y el estar libre de tratamiento degradante (para esto último se puede usar el artículo 5 de la Declaración Universal de los Derechos

Humanos), o valores como el bienestar, la privacidad y el consentimiento, y la responsabilidad. Los Sharkeys enfatizan en el bienestar humano. Para la bioética y usados principalmente en la ética médica. Tom Beauchamp y James Childress han propuesto con cierta fama los principios de autonomía, beneficencia, no maleficencia y justicia[19]. Se suponen que estos principios deben guiar a todo el personal sanitario y se pueden aplicar a la ética de usar robots en la asistencia sanitaria. Estos derechos y principios, sin embargo, están formulados a nivel general y en la práctica están formulados en su mayoría de forma negativa: esto es, como prohibiciones. Aún continúa siendo una cuestión abierta qué tipo de concepción positiva del cuidado deberíamos tener por objetivo. Por ejemplo, los principios bioéticos de no dañar a los pacientes y de hacer el bien ofrecen una guía bastante limitada para cuestiones concretas sobre si los robots se deberían introducir en la asistencia sanitaria y, si así fuera, para qué tareas y roles, y para trabajar con qué tipo de pacientes. Los principios bioéticos también fracasan a la hora de subrayar de manera suficiente la dimensión social de la asistencia sanitaria.

Por el contrario, los Sparrow aceptan en su artículo que una asistencia sanitaria de buena calidad debe incluir cuidado emocional (real) y relaciones sociales, dadas las necesidades sociales y emocionales de las personas[20]. Tanto los Sparrow como los Sharkey mantienen la opinión de que realizar tareas de cuidados rutinarias a menudo proporciona una oportunidad para la interacción social; para los ancianos, pero, por ejemplo, para los niños (ver también el capítulo anterior), y, por tanto, contribuye a una necesidad y derecho humano importante[21].

En línea con esta percepción sobre la dimensión social del cuidado, pero haciendo referencia a otros principios generales y teoría normativa, he propuesto que podemos usar la versión de Martha Nussbaum del enfoque basado en las capacidades para evaluar a los robots en la asistencia sanitaria[22]. El enfoque basado en las capacidades está diseñado para ofrecer una guía sobre el desarrollo humano, pero también es relevante para reflexionar de forma más general sobre la calidad de vida, la dignidad humana y la justicia. La lista de Nussbaum incluye las siguientes «capacidades humanas centrales»:

1. Vida: «ser capaz de vivir hasta el final de una vida humana de duración normal. No morir de forma prematura, o antes de que la

propia vida se vea tan reducida como para que no merezca la pena vivirla»[23].

2. Salud corporal (incluyendo alimentación y alojamiento).

3. Integridad corporal: libertad de movimientos, libertad frente a agresiones sexuales y violencia; tener oportunidades de satisfacción sexual.

4. Ser capaz de utilizar tus sentidos, imaginación y pensamiento: experimentar y producir cultura, así como libertad de expresión y de religión.

5. Emociones: ser capaz de tener relaciones afectivas con cosas y personas.

6. Razón práctica: ser capaz de formarse una concepción de lo bueno y participar de una reflexión crítica sobre la planificación de la vida de cada uno.

7. Afiliación: ser capaz de vivir con y hacia otros, imaginar al otro y respetarlo.

8. Otras especies: ser capaz de vivir con preocupación por los animales, las plantas y la naturaleza.

9. Juego: ser capaz de reír, jugar y disfrutar de actividades recreativas.

10. Control sobre el entorno de cada uno: elección política y participación, ser capaz de tener propiedades, y ser capaz de trabajar como un ser humano con reconocimiento mutuo[24].

Aplicado a la buena asistencia sanitaria, usar estos criterios implica que es necesario, pero no suficiente, si el cuidado ayuda a recuperar la salud de los pacientes o a mantener vivos a los ancianos: también debería contribuir a su bienestar social y emocional, conducir al respeto hacia los demás, otorgar a las personas control sobre su entorno, tener en cuenta los planes propios de cada persona, y así sucesivamente. Entonces, la cuestión relativa a los robots debería discutirse en el contexto de una concepción más amplia de la asistencia sanitaria y estar dirigida al aumento, mantenimiento y recuperación de las capacidades humanas centrales. Por ejemplo, ¿realmente, mejora o mantiene la tecnología la capacidad de afiliación con otros, o disminuye el contacto social, como sospechan los Sparrow? ¿Respeta la introducción de robots asistenciales la capacidad de las personas para la razón práctica o, por ejemplo, las infantiliza, amenazando, así, su dignidad (por ejemplo, de las

personas mayores)? Y si consideramos la cuestión desde el punto de vista de la justicia, ¿estarán solo disponibles estos beneficios para los ricos en sociedades avanzadas tecnológicamente?[25]. A pesar de que quizás no se debería pedir *demasiado* al personal sanitario, cuestiones como estas pueden ayudarnos a reflexionar sobre qué tipos de cuidados queremos, cuál es a su vez un buen punto de anclaje ético para plantear cuestiones sobre los robots asistenciales. Este enfoque ha inspirado a otros expertos a proponer un enfoque de las capacidades en relación al cuidado robótico[26]. El enfoque de las capacidades puede también ofrecer una forma de ampliar las definiciones de salud y bienestar, desde la salud corporal hasta las dimensiones mentales y sociales de la salud y, preferiblemente, ir más allá totalmente de las categorías dualistas. Esto es importante para la ética del cuidado robótico y del cuidado en general.

Otro enfoque, que apunta hacia la cuestión del buen cuidado poniéndolo en el contexto de la calidad de las vidas humanas en general, es hacer la antigua pregunta filosófica aristotélica sobre *la vida buena (eudaimonia)* y derivar alguna guía normativa desde ella. Por ejemplo, ¿qué significa el florecimiento humano para la gente enferma o anciana? ¿Podría la noción de Aristóteles de amistad guiar él cuidado por parte de familiares y amigos, o es demasiado exigente, y diseñada para relaciones entre iguales, inadecuada para reflexionar sobre los cuidadores y los dependientes? ¿Qué significa la compañía en el contexto de la asistencia sanitaria? ¿Qué es la asistencia sanitaria virtuosa, y cuál podría ser el papel de los robots en ese tipo de cuidado?[27]. Sparrow también se refiere a la virtud aristotélica y a la buena vida cuando defiende el cuidado de los mayores que cumple con el requisito ético del respeto, así como el «bien objetivo» del reconocimiento[28].

En una línea relacionada, he preguntado qué es el *trabajo* virtuoso del cuidado[29]. He defendido que tal trabajo no es solo social, sino que implica también un compromiso cualificado y corporal: un tipo de artesanía personal (*craftspersonship*) que depende del conocimiento tácito y del saber-cómo. Además, está dirigido hacia la excelencia o la virtud[30]. Los cuidadores son y deberían ser artesanos personales virtuosos, si y en la medida en que se comprometen de forma diestra, corporal y socialmente con las personas y las cosas para alcanzar la excelencia en el cuidado. Por ejemplo, se puede considerar que las enfermeras hacen exactamente este tipo de trabajo (o, al menos, deberían estar haciendo este tipo de trabajo), cuando se comprometen

con los pacientes y con todo tipo de cosas de una forma competente, que implica un saber-cómo que está muy próximo al «saber-qué» (saber teórico). ¿Pueden alcanzar los robots tal artesanía personal? ¿O deberían usarse solo para ayudar a los humanos en la artesanía de los cuidados?

Basándose en el trabajo de Stuart y Hubert Dreyfus, uno se podría preguntar si la *pericia*, del cuidado y el saber-cómo ético relacionado, los cuales implican ambos un tipo de saber-cómo encarnado e intuitivo, se pueden transferir a los robots[31]. Considérese de nuevo la cirugía robótica. Aquí el robot está siendo manejado aún de forma remota por el cirujano, que tiene que adquirir unas destrezas particulares y una pericia encarnada para teleoperar el robot. En su trabajo (y como parte de él), entendido como un tipo de artesanía, el cirujano habrá desarrollado un sentido de lo que es éticamente aceptable y una buena práctica. En este sentido, se retiene la artesanía o, al menos, se aprende un nuevo tipo de ella: el oficio de operar mediante el robot. Pero ¿puede el robot asumir la tarea en su totalidad, alcanzando la pericia técnica y ética del humano? Estas consideraciones sobre la asistencia sanitaria y la pericia del cuidado plantean la cuestión de si la asistencia sanitaria, entendida como artesanía y pericia, puede ser realizada por robots[32]. Existe al menos el peligro de que la interacción habilidosa y la artesanía se vean erosionadas cuando el robot sustituya al personal sanitario humano.

Sin embargo, un problema tanto con el enfoque de las capacidades como con el de la vida buena es que su énfasis aristotélico sobre el florecimiento humano puede considerarse que privilegia aquellas fases de la vida en las que somos adultos, independientes, no estamos enfermos, no somos ancianos, y tenemos nuestras capacidades completas. Puede cuestionarse hasta qué punto es esto adecuado para otras fases de la vida cuando hay un «declive natural en las capacidades»[33]. Como respuesta, se podría adaptar el enfoque de la vida buena (por ejemplo, hablo sobre «la mejor vida posible») y cuestionar la imagen de las personas mayores como personas en declive, o usar otro enfoque[34].

Basándome en mi revisión de la discusión sobre los robots en la asistencia sanitaria y para fines más prácticos, he propuesto algunos criterios generales para el buen cuidado que responden de forma directa a algunas de las cuestiones más importantes identificadas en la sección previa e incluyen algunos de los principios generales esbozados en esta sección, pero sin referencias directas al marco aristotélico o al ideal de la artesanía. Lo que

llamo «un ideal normativo para el buen cuidado» incluye los siguientes criterios de trabajo:

- El buen cuidado intenta recuperar, mantener y mejorar la salud de las personas.
- El buen cuidado se practica dentro de los límites establecidos por los principios bioéticos y códigos de ética profesionales.
- El buen cuidado implica una importante cantidad de contacto humano.
- El buen cuidado no significa solo cuidado físico, sino que tiene también dimensiones psicológicas y relaciónales, tales como la dimensión emocional.
- El buen cuidado no es solo cuidado profesional, sino que también debería involucrar a familiares, amigos y amistades próximos hasta un grado significativo.
- El buen cuidado no (solo) se experimenta como una carga, sino que puede también ser (experimentado como) significativo y valioso.
- El buen cuidado implica compromiso de los expertos con los pacientes (saber-cómo), junto con formas de pericia más formales (saber-qué).
- El buen cuidado requiere un contexto organizativo en el que existen límites a la división del trabajo para no hacer que los criterios previos sean imposibles de cumplir.
- El buen cuidado implica un contexto organizativo en el que las consideraciones financieros-económicas no son los únicos o, incluso, el principal criterio en la organización del cuidado.
- El buen cuidado requiere que el paciente acepte algún grado de vulnerabilidad y dependencia de los demás[35].

Este último criterio concierne a las personas que reciben los cuidados; yendo más allá de las discusiones centradas sobre el personal sanitario y sus obligaciones, el supuesto es que el buen cuidado no es solo una cuestión de lo que ese personal hace, sino que también depende de los que reciben el cuidado, que comparten, al menos, parte de la responsabilidad del buen cuidado (por ejemplo, el autocuidado) y que tienen (deberían tener) diferentes opciones para lidiar con las nuevas tecnologías.

Por supuesto que estos criterios para (objetivamente) un buen cuidado también pueden y deben usarse para evaluar la atención sanitaria sin robots. Además, en mi opinión, estas listas deberían usarse junto con las experiencias

subjetivas de la gente, que es, ciertamente, lo único que importa (ver de nuevo el experimento mental), y que debería, al menos, importar *también*. Al igual que no es bueno tener un buen cuidado subjetivo sin beneficios objetivos, sería equivocado tener expertos diciendo a la gente que reciben un buen cuidado sin que *sientan* que es bueno. Ambas dimensiones son necesarias para el buen cuidado.

Este criterio subjetivo, junto con la exigencia de respetar la dignidad humana, nos recuerda que no se debería olvidar *preguntar* de forma efectiva a la gente implicada: al personal sanitario, a los receptores del cuidado y a otras partes interesadas. Por ejemplo, en el caso de los robots para niños autistas, hemos realizado una encuesta y, en una fase posterior, organizado un taller con las partes interesadas[36]. La razón fundamental es que, en lugar de dejar las decisiones solo en manos de los expertos, sería mejor tener un proceso deliberativo que dé voz a la gente sobre cómo se usa la tecnología (y, preferiblemente, si se debería usar). Por ejemplo, se podría implicar a los ancianos para comprobar la ética del uso de los robots en el cuidado de ancianos[37]. Este enfoque también encaja con el concepto de investigación e innovación responsables, que concierne al menos a dos aspectos: uno sobre hacer ética al principio del proceso de desarrollo e innovación en lugar de después, y otro sobre implicar a las partes interesadas.

El primer aspecto es importante, ya que los ingenieros, los informáticos y otras personas implicadas en el desarrollo de los robots deberían tener corresponsabilidad por su uso y consecuencias. A veces se formula este enfoque en términos de valores. Desarrollado por Batya Friedman y otros, el llamado diseño sensible al valor toma en consideración los valores humanos a lo largo del proceso de diseño[38]. El concepto se ha aplicado también a la robótica. Aimee van Wynsberghe, por ejemplo, ha defendido que la ética debería estar incluida en el proceso de diseño de robots asistenciales, promoviendo los valores y la dignidad de los pacientes[39].

El segundo aspecto subraya la corresponsabilidad de otras partes interesadas y se relaciona a menudo con las ideas del diseño participativo en tanto que ligado a la democracia participativa. Trata de cerrar la brecha entre la tecnología y la sociedad:

La investigación y la innovación responsables son procesos transparentes e interactivos mediante los cuales los actores e innovadores sociales se vuelven mutuamente responsables con una perspectiva sobre la aceptabilidad (ética), la sostenibilidad y la deseabilidad social del proceso de innovación y

sus productos comercializables (para permitir una adecuada integración de los avances científicos y tecnológicos en nuestra sociedad)[40].

> El buen cuidado no significa solo cuidado físico, sino que tiene también dimensiones psicológicas y relacionales, tales como la dimensión emocional.

De nuevo, este concepto se puede aplicar, y se ha aplicado, a la robótica. A la luz de esta concepción de la investigación responsable y de la innovación, Bernd Stahl y yo mismo hemos defendido una ética robótica del buen cuidado sanitario que, junto con un análisis de los problemas éticos, incluye formas de reflexión, diálogo y experimentación que se acercan a las prácticas de innovación y a los contextos de uso[41]. Este es otro significado del «buen cuidado» y, se puede defender, de la buena robótica y la buena *tecnología*.

En conclusión: reflexionar sobre los robots como asistentes para los cuidados no solo aborda problemas éticos específicos planteados por los robots asistenciales, sino que también nos hace reflexionar sobre lo que es el buen cuidado, y cómo asegurarnos de que los valores relevantes y las partes interesadas estén integrados en el área de los cuidados junto con el desarrollo y el uso de la tecnología pertinente.

Capítulo 5

Coches autónomos, agencia moral y responsabilidad

Imaginemos que un coche autónomo conduce a alta velocidad a través de una calle estrecha. Hay niños jugando en la calle. El coche tiene dos opciones: o bien evita a los niños y se estrella contra un muro, probablemente matando al único pasajero humano, o bien continúa su trayectoria y frena, pero probablemente demasiado tarde como para salvar la vida de los niños. ¿Qué debería hacer el coche? ¿Qué harán los coches? ¿Cómo deberían estar programados?

Este experimento mental es un ejemplo del denominado dilema del tranvía. Un tranvía está a punto de atropellar a cinco personas que están atadas a la vía. Estás junto a la vía y puedes tirar de una palanca que redirige al tranvía hacia otra vía, donde hay atada una persona. ¿Tiras de la palanca? Si no haces nada, morirán cinco personas. Si tiras de la palanca, morirá una persona. Este tipo de dilema se usa a menudo para hacer que la gente piense sobre los dilemas morales planteados por los coches autónomos. La idea es que tales datos podrían ayudar a las máquinas a decidir. Por ejemplo, la plataforma *online* Moral Machine (máquina moral) ha reunido millones de decisiones de usuarios de todo el mundo sobre sus preferencias morales en casos en los que un conductor debe escoger «el menor de dos males[1]». Se les preguntaba a las personas si un coche autónomo debería priorizar a los humanos sobre las mascotas, a los pasajeros sobre los peatones, a las mujeres sobre los hombres, y así sucesivamente. Resultaron interesantes las diferencias culturales relativas a las elecciones realizadas[2]. Algunas culturas (colectivistas) como, pongamos por caso, Japón y China, eran menos propensas a salvar a los jóvenes en vez de a los mayores, mientras que otras culturas (individualistas), como las de Reino Unido o los Estados Unidos, eran más propensas a salvar a los jóvenes. Así, este experimento no ofrece solamente una forma de acercarse a la ética de las máquinas, sino que también

plantea la cuestión más general sobre cómo tener en cuenta las diferencias culturales en la robótica y en la automatización.

La figura 3 muestra un ejemplo de una situación en la que se da el dilema del tranvía: ¿debería continuar el coche su camino y matar a cinco peatones, o debería cambiar su rumbo y matar a uno?

Aplicar el dilema del tranvía al caso de los coches autónomos puede no ser la mejor forma de reflexionar sobre la ética de los coches autónomos. Por suerte, raramente nos encontramos con situaciones así en el tráfico, o los desafíos pueden ser más complicados y no implicar elecciones binarias, y esta definición del problema refleja un enfoque normativo específico hacia la ética (hacia el consecuencialismo y, en particular, hacia el utilitarismo). Se discute en los escritos sobre el tema hasta qué punto los dilemas del tranvía representan los desafíos éticos reales[3]. Sin embargo, los dilemas del tranvía se usan a menudo como una ilustración de la idea de que cuando los robots se vuelvan más autónomos, tendremos que reflexionar sobre si vamos o no a otorgarles algún tipo de moralidad (si es que se puede evitar hacerlo), y, si así fuera, qué tipo de moralidad.

Además, los robots autónomos plantean cuestiones relacionadas con la responsabilidad moral. Considérese de nuevo al coche autónomo. En marzo de 2018, un coche autónomo de Uber mató a un peatón en Tempe, Arizona. Había un operador en el coche, pero en el momento del accidente el coche estaba en modo autónomo. El peatón estaba caminando fuera del paso de cebra. El SUV Volvo no frenó a medida que se aproximaba a la mujer[4]. Este no es el único caso fatal del que se ha dado cuenta. En 2016, por ejemplo, un modelo S de Tesla en modo piloto automático falló a la hora de detectar un camión con un tráiler grandes que estaban cruzando la autopista e impactó con él, matando al conductor del Tesla. Para muchos observadores, tales accidentes muestran no solo las limitaciones del desarrollo tecnológico actual (a día de hoy no parece que los coches autónomos estén listos para salir a la carretera) y la necesidad de regulación; plantean desafíos con respecto a la atribución de responsabilidad. Considérese el caso de Uber. ¿Quién es responsable del accidente? El coche no puede asumir la responsabilidad. Pero las partes humanas implicadas pueden ser todas potencialmente responsables: la compañía Uber, que emplea un coche que no está aún listo para salir a la carretera; el fabricante Volvo, que fracasó a la hora de desarrollar un coche seguro; el operador que estaba en el coche, que no reaccionó a tiempo para parar el vehículo; el peatón, que no estaba caminando por el paso de cebra;

los reguladores (esto es, el estado de Arizona), que permitió que este coche, se testara en la carretera. ¿Cómo debemos atribuir y distribuir la responsabilidad, dado que el coche estaba conduciendo de forma autónoma y había tantas partes involucradas? ¿Cómo distribuiremos la responsabilidad en todos los tipos de casos que involucran robots autónomos, y cómo vamos a lidiar con esta cuestión como profesionales (por ejemplo, como ingenieros), como empresa y como sociedad, preferiblemente de forma proactiva, antes de que ocurran los accidentes?



Figura 3. Ejemplo de un dilema del tranvía presentado a los participantes de Moral Machine.

ALGUNAS CUESTIONES CON RESPECTO A LOS ROBOTS AUTÓNOMOS

Como muestra el accidente de Uber, los coches autónomos no son completamente ciencia ficción. Están siendo probados en la carretera, y los fabricantes de coches los están desarrollando. Por ejemplo, Tesla, BMW y Mercedes ya han probado vehículos autónomos. Muchos de estos coches no son aún completamente autónomos, pero las cosas están yendo en esa dirección. Los coches no son los únicos robots autónomos e inteligentes que tenemos alrededor. Considérese de nuevo los robots autónomos en los hogares

y en los hospitales. ¿Qué pasa si causan daño a las personas? ¿Cómo se puede evitar? ¿Deberían proteger activamente a los humanos del daño? ¿Qué pasa si tienen que realizar elecciones éticas? ¿Tienen la capacidad de tomar tales decisiones? Además, algunos robots están desarrollados para matar (véase el capítulo 7 sobre robots militares). Si escogen su objetivo de forma autónoma, ¿lo podrían hacer de una forma ética (suponiendo, por mor del argumento, que permitimos que esos robots maten)? ¿Qué tipo de ética deberían usar? ¿Los robots pueden siquiera tener ética? Con respecto a los robots autónomos en general, la cuestión es si necesitan algún tipo de moralidad, y si es esto posible (si podemos y debemos tener «máquinas morales»). ¿Pueden tener agencia moral? ¿Qué es la agencia moral? ¿Pueden los robots ser responsables? ¿Quién o qué es y debe ser responsable si algo sale mal?

¿Máquinas morales? ¿Pueden los robots ser agentes morales?

Comencemos con la cuestión relacionada con la agencia moral y las máquinas morales. Algunos investigadores piensan que los robots pueden ser agentes morales e intentan desarrollar máquinas morales. Por ejemplo, Michael y Susan Anderson hacen «máquinas éticas» en el sentido de que buscan crear agentes éticos que tomen decisiones de la forma en que lo hacen los humanos. Creen que «la ética puede hacerse computable», y que debemos otorgar a las máquinas principios y dejarlas razonar de una forma racional, evitando, por lo tanto, que los seres humanos se «dejen llevar por sus emociones[5]». Alan Winfield, un robotista, también cree que podemos crear máquinas morales, que define en una de sus charlas como «un robot capaz de decidir o moderar sus acciones sobre la base de reglas éticas[6]». ¿Qué tipo de reglas? Winfield ofrece el ejemplo de un robot que evita que un humano caiga en un hoyo al seguir la regla de que los humanos no deberían recibir daño. Winfield se refiere a la primera ley de las leyes de la robótica de Isaac Asimov, que este propuso en sus historias de ciencia ficción. La primera ley es que «un robot no puede causar daño a un ser humano o, por inacción, permitir que un ser humano sufra daño[7]». En términos de, teoría moral normativa, el robot seguiría una ética consecuencialista, evitando malas consecuencias para los humanos (por ejemplo, el daño de alguien que cae en un hoyo). Para este propósito, argumenta Winfield, no necesitamos un robot sintiente, sino más bien uno que siga reglas éticas, y para ser capaces de implementar esas reglas, debe ser capaz de tener un modelo de sí mismo y de

su entorno (incluyendo a los otros en ese entorno) para predecir las consecuencias de sus propias acciones, así como las de los demás. Un robot con un «modelo interno» así sería capaz de generar hipótesis qué-si como, por ejemplo: ¿qué pasa si llevo a cabo la acción *x*, y cuál de las siguientes posibles acciones subsiguientes debo escoger[8]? En el ejemplo del hoyo, el robot necesita ser capaz de predecir lo que va a hacer el humano (caer en un hoyo) y averiguar qué tiene que hacer para prevenirlo. La razón principal para desarrollar tales robots éticos es que una vez que tengamos robots más autónomos e inteligentes que puedan asumir todo tipo de tareas y responder al entorno, los robots seguros dejarán de ser suficientes. Como escriben Wendell Wallach y Colin Allen en su libro *Moral Machines* (Máquinas morales),

> Si se ha de confiar en las máquinas multipropósito, que operan sin ataduras de sus diseñadores o propietarios, y se programan para responder de forma flexible en entornos reales o virtuales, debe haber confianza en que su comportamiento satisfaga las normas apropiadas. Esto va más allá de la seguridad del producto tradicional […] Si un sistema autónomo debe minimizar el daño, debe ser también «consciente» de las posibles consecuencias dañinas de sus acciones, y debe elegir sus acciones a la luz de este «conocimiento», incluso si tales términos se aplican solo metafóricamente a las máquinas[9].

Aun así, no todos creen que la moralidad de los robots autónomos es una cuestión de otorgarles la capacidad de tomar decisiones morales, de crear máquinas morales. Se podría adoptar más bien el punto de vista de que los robots tengan consecuencias morales, pero que los humanos deberían realizar las evaluaciones éticas. Se podría defender que los robots no pueden ser de ninguna manera agentes morales porque carecen de consciencia y de otras precondiciones para la agencia moral, que la ética no puede ni debe reducirse a seguir reglas y principios, o que tener capacidades como la sensibilidad y las emociones es necesario para tomar buenas decisiones éticas y que, por lo tanto, es peligroso querer desarrollar máquinas morales que carezcan de estas capacidades. Visto desde esta perspectiva, nos deberíamos centrar en la ética de los humanos (humanos que usan robots) en lugar de intentar desarrollar máquinas morales (se podrían desarrollar argumentos similares contra lo que se podría llamar «máquinas políticas»: dejemos a los humanos evaluar el impacto político de los robots, como el impacto sobre la justicia).

Por ejemplo, Deborah Johnson ha argumentado que dado que los sistemas computacionales (y, por tanto, los robots) no tienen estados mentales e «intenciones» como agentes libres, no cumplen una condición clave para la

agencia moral[10]. Más bien, son componentes de la acción humana, particularmente del diseñador y el usuario.

La ética se debería centrar en la acción humana. Además, los robots (carecen de capacidades emocionales, que son indispensables para el juicio moral. Sin esta capacidad, los robots que siguen reglas se convertirían en peligrosos robots «psicópatas[11]». Incluso aparte de estas emociones, como muestran las propias historias de Asimov, los principios y las leyes a menudo chocan entre sí y crean dilemas difíciles[12]. Como señala Nick Bostrom, los principios de Asimov probablemente estuvieran formulados para que «fracasaran de formas interesantes»: para hacer una trama interesante[13].

Los robots carecen de capacidades emocionales, que son indispensables para el juicio moral. Sin esta capacidad, los robots que siguen reglas se convertirían en peligrosos robots «psicópatas».

Pero si estas objeciones son válidas y los robots no pueden ser agentes morales *completos*, ¿pueden considerarse agentes morales en algún grado y para propósitos prácticos? Esta es una especie de posición intermedia adoptada por Wallach y Allen. Mantienen que los robots pueden y deben tener «moralidad funcional», pero no afirman que esta sea una agencia moral completa. Esta última bien podrá pertenecer siempre a la ciencia ficción, pero podemos desarrollar sistemas que tengan «alguna capacidad para evaluar las ramificaciones éticas de sus acciones» y, en este sentido, «la capacidad para evaluar y responder a desafíos morales», incluso si no son agentes morales completos. Wallach y Allen ofrecen el ejemplo de un piloto automático que pueda disponer de sensibilidad ética[14]. Podría interpretarse también que los

robots de Winfiled tienen algún grado de moralidad funcional, pero no moralidad completa. Dicho esto, el término «máquinas morales» continúa siendo problemático, ya que sugiere la posesión completa de tales capacidades.

Otro enfoque consiste en rebajar los criterios para la agencia moral y distanciarse de la suposición de que la agencia moral de los robots debería modelarse sobre la moralidad humana. Luciano Floridi y J. W. Sanders han argumentado que los agentes morales no tienen necesariamente que exhibir estados mentales: una «moralidad no mental» puede ser suficiente[15]. Si en un nivel de abstracción escogido por nosotros existe un grado suficiente de interactividad, autonomía y adaptabilidad, y si el agente es capaz de acciones calificables como morales, entonces es un agente moral. Por ejemplo, un perro de búsqueda y rescate no tiene una mente como los humanos, sin embargo, es un agente moral. Basándonos en este razonamiento, parece que un robot puede tener una moralidad no mental si al ser un agente (dado que tiene un grado suficiente de interactividad, autonomía y adaptabilidad en algún nivel de abstracción), es capaz de acciones calificables como morales. Inspirado por esta explicación, John Sullins ha defendido que no es necesario que los robots tengan personalidad para ser calificados como agentes morales: basta con que el robot sea significativamente autónomo para los humanos, que se puede analizar o explicar el comportamiento del robot adscribiéndole alguna intención de hacer el bien o causar daño, y que el robot se comporte de una forma que muestre cierta comprensión de su responsabilidad hacia algún otro agente moral. Por ejemplo, si un robot asistencial lleva a cabo algunas tareas como los humanos, y lo hace de forma autónoma, se comporta de maneras en las que percibimos que actúa de una forma intencional, y que parece que comprende su papel y responsabilidad en el contexto de la asistencia sanitaria, entonces la máquina es un agente moral completo[16]. Se podría objetar que este debilitamiento de los criterios es problemático y peligroso. Al construir el robot como un agente moral, esperamos que exhiba de su parte un comportamiento moral que sea similar por completo al humano y confiamos en él. Pero esto es un error, ya que estamos siendo engañados en cuanto a las capacidades del robot: el robot nunca cumplirá estas expectativas debido a las razones ofrecidas anteriormente, como que no tiene intenciones reales, no tiene emociones, no es consciente, etc.

Sullins conecta la agencia moral con la responsabilidad. Pero ¿pueden los robots ser responsables? Como comenta acertadamente Sven Nuholm, incluso

si se cree que los principios morales necesitan estar programados en los robots, «esto no resuelve aún la cuestión de quién es responsable si y cuando la gente resulta herida o muerta por la acción de sistemas autónomos»[17]. De forma más general, ¿cómo vamos a lidiar con la responsabilidad cuando robots más autónomos e inteligentes asuman las tareas de los humanos? En la siguiente sección discutiré la cuestión de la responsabilidad suponiendo a la vez que los robots no pueden ser agentes morales completos y (por tanto) no pueden ser responsables.

LA ATRIBUCIÓN DE RESPONSABILIDAD EN EL CASO DE LOS ROBOTS AUTÓNOMOS

A pesar de que se puede ser responsable no solo por las malas acciones y sus consecuencias, sino también por las buenas, normalmente se plantea la cuestión de la atribución cuando algo sale mal o, mirando hacia el futuro, intentamos evitar que algo vaya mal. Este también es el caso en la ética de los robots. Piénsese de nuevo en los casos de accidentes con coches autónomos. Se pueden plantear cuestiones sobre cómo atribuir la responsabilidad cuando algo sale mal, dado que el usuario no está conduciendo realmente y no puede intervenir[18].

Los robots más *autónomos* plantean especialmente esta cuestión con respecto a la atribución de responsabilidad. Normalmente responsabilizamos a las personas de lo que hacen. Pero ¿qué pasa si el robot, como en el caso de un coche, asume las tareas de los humanos? Si suponemos que el robot no es ni puede ser un agente moral completo, entonces se sigue que el robot no puede ser responsable. Sin embargo, *tiene* agencia. El aumento de la autonomía y de la agencia del robot, pero no de su consiguiente responsabilidad, crea lo que Andreas Mathias llama «brecha de responsabilidad[19]». ¿Quién *es* responsable si no se puede responsabilizar al robot, y si el usuario no es capaz (o se supone que no lo es) de intervenir? ¿Y qué se puede esperar razonablemente de los usuarios (por ejemplo, conductores de coches) en términos de conocimiento sobre el sistema técnico y los riesgos de utilizarlo? ¿Qué pasa si el comportamiento de la máquina no se puede predecir?

En respuesta a este problema, podríamos insistir que los humanos (usuarios y otras personas) deberían asumir la responsabilidad de lo que hace el robot autónomo. Pero esto plantea otra cuestión: ¿bajo qué condiciones *pueden* hacerse responsables? Desde Aristóteles, los filósofos han definido al

menos dos condiciones tradicionales para atribuir responsabilidad por una acción: la llamada condición de control y la condición «epistémica»[20]. Eres responsable si eres el agente de la acción y tienes un grado de control suficiente sobre ella, y si sabes y eres consciente de lo que estás haciendo. Aristóteles defendió en la *Ética a Nicómaco* que la acción debe tener su origen en el agente, y uno no puede ser ignorante de lo que está haciendo[21]. ¿Se cumplen estas condiciones en el caso del uso de los robots autónomos?

Comencemos con la condición de agencia (voluntaria) y la de control. Existen varios problemas en el caso de los robots autónomos.

En primer lugar, como indica el problema de la brecha de responsabilidad, en el caso de los robots autónomos los humanos pueden no tener un control directo o total sobre ellos. En respuesta, se podría exigir que los humanos deberían tener siempre un control suficiente. Pero ¿qué pasa si no hay tiempo para intervenir, que podría ocurrir en situaciones en las que intervienen sistemas de armas de reacción rápida o el comercio de alta frecuencia? ¿O qué pasa si el sistema puede anular las decisiones humanas, como sucederá en un futuro próximo con los sistemas de piloto automático en aeronaves? En respuesta a esto, se podría exigir que tales robots y sistemas autónomos no se deberían construir. Sin embargo, ¿qué pasa si otros países desarrollan estas armas? ¿Y si resulta que anular las decisiones humanas en el control de la aeronave hace que los viajes aéreos sean más seguros?

En segundo lugar, como en muchas de las acciones tecnológicas, hay normalmente muchas personas implicadas en el desarrollo y el uso de los robots autónomos. Hay «muchas manos»[22]. Esto hace que sea difícil responsabilizar a un individuo (y no a otros), o incluso encontrar *un* individuo que pueda hacerse responsable. ¿Si se implica a más gente, cómo se distribuye la responsabilidad? Pensemos de nuevo en el caso de Tesla, en el que, en principio, mucha gente ha contribuido al accidente. En respuesta a ello, se podría defender un concepto de responsabilidad distribuida, pero cómo distribuir la responsabilidad de forma exacta continúa siendo un problema.

En tercer lugar, existe una dimensión temporal. En el desarrollo, uso y mantenimiento de tecnologías como los robots, a menudo hay una larga cadena de agencia y causas humanas. Puede no estar claro quién hizo qué en qué momento, y qué causó que tal cosa fallara en tal momento. Piénsese en un avión que falla o, de nuevo, en un coche contemporáneo. En el caso de tales sistemas tecnológicos complejos, en los que el robot o la automatización es

una parte, puede ser difícil trazar y distribuir la responsabilidad dada la larga cadena de acciones y causas.

En cuarto lugar, no solo hay muchas manos humanas, sino también «muchas cosas». En la robótica están involucradas muchas tecnologías y componentes distintos, materiales (*hardware*) e inmateriales (*software/código*). Todos estos componentes están unidos a la acción humana (en algún punto en el tiempo). También hay varias relaciones entre las partes, como, por ejemplo, varias interfaces y conexiones. Considérese de nuevo un coche autónomo. En caso de un accidente, puede no estar claro qué componente falló, cómo se causó, y quién es responsable de su fallo.

Finalmente, no está claro hasta qué punto es «voluntario» y «libre» el uso de robots autónomos, dado que el usuario final puede no entender al robot. Por ejemplo, un usuario de un coche autónomo podría no entender los riesgos de usar tal coche (por ejemplo, un taxi), no digamos el sistema técnico que automatiza el coche, y los modelos y suposiciones que están en funcionamiento en el *software*. Esto nos conduce hasta la siguiente condición: el conocimiento.

La segunda condición aristotélica es la epistémica[23]. Para ser responsable, se debe saber lo que se está haciendo o, formulado negativamente por Aristóteles, no se puede ser ignorante. Aristóteles distingue aquí entre varias formas en las que se puede ser ignorante: «un hombre puede ignorar quién es, qué está haciendo, sobre qué o quién está actuando y, a veces, también qué (por ejemplo, el instrumento) está haciendo con ello, y cómo lo está haciendo (por ejemplo, si lo hace con cuidado o de forma violenta)[24]». Un término positivo es la conciencia. Fernando Rudy-Hiller distingue entre diferentes tipos de conciencia que son necesarios: la conciencia de la acción, la conciencia del significado moral de la acción, la conciencia de las consecuencias de la acción, y (de acuerdo con algunos) la conciencia de que hay alternativas[25]. Pero Aristóteles también incluye un elemento que concierne a la ignorancia sobre el *instrumento*, «qué… está haciendo con él», ya que conocer la tecnología que estás utilizando es importante para la responsabilidad.

Para la robótica autónoma, este criterio epistémico implica (entre otras cosas) que los desarrolladores y los usuarios deberían ser informados de modo que fueran conscientes de sus acciones, de sus consecuencias y de la tecnología que están usando. En el caso de los robots autónomos, sin embargo, no siempre es predecible lo que va a hacer el robot y qué

consecuencias involuntarias habrá. Por ejemplo, no estaba claro para los desarrolladores del Tesla del que hablamos que no reconocería al peatón, ni había la intención de que el coche lo matara (y de hecho, que el coche matara a alguien). Además, en el caso de robots con IA, lo que el robot hace puede no ser completamente transparente incluso para sus desarrolladores. Esto es especialmente relevante en el caso de los denominados algoritmos de «caja negra» en aprendizaje automático cuando trabajan con redes neurales: aquí no está claro cómo llegó exactamente el sistema a su decisión. Pero la falta de transparencia y conocimiento también es un problema más general en la IA, la ciencia de datos y la robótica autónoma. La brecha de responsabilidad se relaciona entonces con una *brecha, de conocimiento*: una brecha entre, por una parte, los desarrolladores y el sistema y, por otra, sus usuarios y aquellos que están (efectiva o potencialmente) afectados por el sistema. El que haya tantas manos y tantas cosas, así como relaciones en diferentes momentos, plantea también problemas de conocimiento. Por ejemplo, puede que hayan trabajado muchos desarrolladores en el código del coche. Los últimos desarrolladores (y, por supuesto, los usuarios y otras partes implicadas) pueden no conocer algunos aspectos del código y lo que este hace. Estos tipos de ignorancia son moralmente problemáticos, ya que implican efectivamente que, en los sentidos expuestos, tanto los desarrolladores como los usuarios de robots autónomos *no saben por completo qué están haciendo*. En respuesta a esto, se podría intentar sensibilizar sobre estas cuestiones a los usuarios y desarrolladores, y asegurarse de que hay suficiente conocimiento para colmar la brecha y evitar los correspondientes problemas de ignorancia aristotélicos.

Sin embargo, más allá del pensamiento aristotélico, existe una tercera condición moral de responsabilidad que a menudo no se menciona, pero que es importante añadir una vez que adoptamos un enfoque más relacional. La responsabilidad no es solo un problema que concierne a la agencia y a las condiciones de conocimiento relacionadas. Desde un punto de vista relacional, no deberíamos preguntar solamente quién es responsable y de qué, sino considerar también *ante quién* somos responsables[26]. La responsabilidad no requiere solo que se tome control (voluntario) de la acción y se sepa lo que se está haciendo. Se debería también ser capaz de «responder» ante alguien. Se debería estar dispuesto a explicar *a* alguien lo que se hace y decide, y ser capaz de hacerlo. La responsabilidad trata también de «rendición de cuentas[27]». Esta condición de «responsabilidad como

rendición de cuentas» está relacionada con el control y el conocimiento, pero tiene su propia importancia específica y moral.

Mientras que las condiciones previas solo se centran en la responsabilidad del *agente*, aquí el *paciente* responsable se halla en el centro de la atención ética[28]. Esos pacientes incluyen a aquellos afectados de forma inmediata por la acción, como aquellos sobre los que habla Aristóteles, pero también aquellos que se ven *potencialmente* afectados y otras partes implicadas.

Aplicado al problema de la responsabilidad de robots y sistemas autónomos, esto implica que aquellos que desarrollan y usan robots deberían no solo tener un control suficiente sobre el robot y saber lo que están haciendo; deberían estar dispuestos y ser capaces de responder a preguntas de aquellos que se ven afectados, de forma directa o indirecta, por las acciones del robot. Por ejemplo, los que programan los coches autónomos y las empresas que emplean el coche (por ejemplo, una compañía de taxis) deberían ser capaces de rendir cuentas ante las víctimas (de hecho o potenciales) y a sus familias en accidentes en los que el coche esté involucrado. Como usuarios, los pilotos y las aerolíneas para las que trabajan deberían ser capaces de explicar por qué un piloto automático tomó una decisión específica, tanto en el contexto de un accidente como en una incidencia que podría haber causado un accidente, incluso si el piloto automático depende de una IA. Hablando proactivamente, los desarrolladores y los usuarios de tecnología robótica autónoma deberían desarrollar y usar la tecnología de una forma en la que continúen siendo capaces de rendir cuentas ante aquellos que puedan beneficiarse o sufrir a causa de lo que el robot haga junto con las consecuencias de sus acciones.

Cumplir con esta exigencia de responsabilidad como rendición de cuentas puede ser un reto, dadas las dificultades que plantea la condición de conocimiento, ya que, por ejemplo, las consecuencias potenciales involuntarias pueden ser difíciles de predecir, o los usuarios finales y las empresas pueden no saber realmente lo que están haciendo en el sentido de que emplean una tecnología que no comprenden. En respuesta a esto, se debería subrayar de nuevo la importancia de la consciencia y del conocimiento de la tecnología y, en lugar de simplemente culpar a las personas, pedir a los desarrolladores y a los usuarios que se les apoye para obtener tal consciencia y conocimiento. Se podría también crear una regulación que exigiera que solo se pudieran construir robots y sistemas autónomos que cumplan las condiciones de responsabilidad como rendición

de cuentas de diversas formas, incluyendo tecnología más explicable (por ejemplo, cuando está implicada la IA). Se podría abordar también el problema de las consecuencias involuntarias mediante la creación, el análisis y la discusión de varios escenarios futuros. Por ejemplo, se podrían crear escenarios que representasen cómo se podrían usar los robots autónomos de acuerdo con diferentes concepciones del tráfico, las ciudades y la sociedad. Además, se podría exigir que los afectados por la tecnología, los «pacientes responsables», estén incluidos de alguna forma en los procesos de desarrollo y toma de decisiones sobre el uso de la tecnología. Esta exigencia estaría en línea con ideas como la innovación responsable (ver el capítulo anterior). En el caso de los coches autónomos, por ejemplo, se podría preguntar a los taxistas, a los peatones, a los ciclistas, etc., qué piensan sobre la introducción de estos coches, y discutir entonces con ellos diferentes escenarios y cuestiones éticas potenciales.

> Desde un punto de vista relacional no deberíamos preguntar solamente quién es responsable y de qué, sino considerar también *ante quién* somos responsables.

Nótese que la discusión sobre la responsabilidad no debería limitarse a la responsabilidad individual. Se debería plantear como mínimo si puede haber una «responsabilidad colectiva» de los robots autónomos en la sociedad. La cuestión de la responsabilidad relativa a los robots debería incluir, pero no reducirse a, la responsabilidad individual de los ingenieros y de los informáticos, así como la de los políticos. En la medida en que todos nos beneficiamos, así como tenemos algo que decir y arriesgar en el futuro de la tecnología, existe un sentido en el que todos somos responsables de ella. Apelar a la acción individual puede no ser suficiente para realizar los cambios

necesarios: la acción colectiva puede ser también necesaria. Además, la categoría de los pacientes responsables no tiene por qué estar limitada a los humanos. Se puede también incluir a ciertos animales y al medio natural. El futuro de la robótica y los sistemas autónomos afecta al futuro de la sociedad, a otros seres vivos y, quizás, al planeta. Finalmente, a pesar de que es importante atribuir responsabilidad y mejorar la calidad ética de la tecnología robótica, la responsabilidad no tiene por qué ser absoluta, y sus condiciones bien pueden no llegar nunca a cumplirse por completo (o no en todos los casos). También existe un aspecto trágico con respecto a cuestiones que conciernen a la responsabilidad: existencialmente hablando, siempre habrá cosas que ocurran fuera de nuestro control, y que tenemos que aceptar y vivir con ello de alguna manera: como individuos y como sociedad[29].

La responsabilidad distribuida dentro de redes de responsabilidad

Este capítulo ha ofrecido una concepción general sobre dos discusiones con respecto a los robots autónomos: si tienen (pueden tener) agencia moral, y qué significa para los robots autónomos que sean responsables, suponiendo que ellos mismos carecen de agencia moral y, por tanto, de la capacidad para la responsabilidad. Sin embargo, se ha cuestionado esta suposición. Ya mencioné la opinión de Sullin en la primera sección. Pero existen otras respuestas a la cuestión de si los robots tienen las propiedades necesarias para la agencia moral y la responsabilidad.

En primer lugar, inspirado por la teoría contemporánea de los estudios de la ciencia y la tecnología (en particular la teoría actor-red y la obra de Bruno Latour), así como, quizás, por un punto de vista poshumanista (ver el capítulo final), se podría argumentar que, a pesar de que los sistemas artificiales como los coches autónomos no tienen una responsabilidad moral completa, la responsabilidad se distribuye o comparte dentro de una red de agentes humanos y no humanos. Por ejemplo, Wulf y Janina Loh defienden que, en el caso de los coches autónomos, la responsabilidad se encuentra distribuida dentro de una red que incluye a los ingenieros, al operador/conductor y al propio sistema artificial. Además, sostienen que el coche y el operador/conductor, a pesar de ser agentes categóricamente diferentes, constituyen «un sistema híbrido que puede asumir una responsabilidad compartida[30]». Siempre y cuando haya en el coche «algo parecido a un conductor», los autores proponen distribuir la responsabilidad como sigue: el coche es responsable de mantener seguras las operaciones de conducción

estándar basándose en las reglas de circulación, mientras que los humanos son responsables de tomar las decisiones moralmente relevantes no cubiertas por estas reglas. Por ejemplo, dado que los humanos tienen autonomía moral y personal, deberían lidiar con los dilemas morales, y tendrían que hacerlo. No queda claro, sin embargo, en qué sentido exactamente podría el coche ser moralmente responsable, dado que carece de autonomía moral y personal, y que seguir las reglas no es hacer algo moral de acuerdo con las teorías morales que subrayan la autonomía (en particular, las teorías kantianas). También es dudoso si en la práctica se podría siempre distinguir (fácilmente) entre seguir las reglas de circulación y lidiar con otras circunstancias, presumiblemente excepcionales o inesperadas en las que las reglas chocan entre sí (como en el caso de las reglas de Asimov) o no ofrecen la guía suficiente.

En segundo lugar, podría ser que en el futuro los robots parezcan *como si* fueran agentes responsables, exhibiendo alguna responsabilidad «virtual[31]». Después de todo, cuando tratamos con humanos, no podemos entrar a mirar en sus mentes. Confiamos en las apariencias y, sobre esta base, suponemos que son (y deberían ser) agentes morales responsables. Quizás la apariencia es la única cosa en la que podamos confiar. Considérense de nuevo las opiniones de Sullin sobre que la percepción de la intención y la apariencia de comprender la responsabilidad son suficientes[32]. ¿Qué pasaría si los robots pudieran *representar* (*perform*) responsabilidad? ¿Qué pasaría si fueran capaces de aparentar control voluntario, conocimiento y rendición de cuentas? ¿Les otorgaríamos no solo agencia, sino también responsabilidad? ¿Es la apariencia suficiente, o realmente necesitan ser agentes morales? ¿Qué queremos decir por «real» en este caso?

Sin embargo, por ahora, este caso continúa siendo ciencia ficción o un experimento mental filosófico. Los robots que tenemos actualmente no pueden engañarnos *así*. Además, incluso *si* los robots contasen como agentes responsables en algún sentido, esa circunstancia no disminuiría necesariamente la responsabilidad de otros agentes responsables, incluyendo los agentes responsables humanos. La responsabilidad no tiene por qué ser un juego de suma cero. Supuestamente *también* querríamos atribuir responsabilidad a los humanos. Pero ¿responsables de qué? Hay al menos dos posibilidades. Bien se usa de nuevo el concepto de responsabilidad distribuida o compartida, y se asigna una tarea especifica a los humanos. En ese caso, los humanos y los agentes artificiales como los robots serían entonces tratados al

mismo nivel cuando nos referimos a tareas y responsabilidades. O bien se establece una responsabilidad jerárquica y se les otorga a los humanos la responsabilidad general. Se puede usar un modelo de responsabilidad delegada o supervisada para hacer que esto funcione: al final los humanos retendrían entonces toda la responsabilidad, incluso si se delegan tareas particulares al robot y son, en ese sentido, su «responsabilidad». Sin embargo, sigue siendo confuso si lo último es lo que queremos decir cuando hablamos de responsabilidad (moral). Esto abre el debate de nuevo. En cualquier caso, bajo las circunstancias actuales, podemos concluir que sería altamente irresponsable no hacer responsables a los humanos de las acciones y consecuencias de los artefactos más bien limitados y, a veces, graciosamente torpes, que llamamos «robots», tanto automatizados como no.

> ¿Qué pasaría si los robots pudieran *representar* (*perform*) responsabilidad. ¿Qué pasará si fueran capaces de aparentar control voluntario, conocimiento y rendición de cuentas?

Y dados los problemas de responsabilidad indicados en este capítulo, parece acertado evitar desarrollar y emplear por doquier robots y sistemas completamente automatizados que ya no permitan a los humanos controlar, saber y responder por lo que hagan las máquinas.

La cuestión de las apariencias y la virtualidad reaparece en el siguiente capítulo, que no trata sobre nuestra responsabilidad hacia otros humanos, sino más bien hacia los *robots*. Si los robots se parecen a los humanos o, al menos, parecen ser más que máquinas, ¿*les* debemos algo? ¿Pueden ser *pacientes* morales?

Capítulo 6

Androides inquietantes, apariencia y paciencia moral

Ella dijo…
¡No me apagues!
No estoy dormida
Estaré contigo
Si te parece bien…

Enciéndeme,
Pero no me apagues

No puedo decir si ella es real
No puedo confiar en lo que siento
¿Está viva o es solo un sueño?
¿O solo una máquina?

Arjen A. Lucassen,
«Don't Switch Me Off» (*No me apagues*)

Las historias y películas de ciencia ficción como *Blade Runner* y *Westworld* muestran robots que se parecen y comportan como humanos. En el futuro, los robots podrían parecerse cada vez más a los humanos y comportarse como si estuvieran vivos. Con una IA, también mejorarán a la hora de hablar con nosotros. Esto plantea cuestiones descriptivo-hermenéuticas y normativas, cuestiones sobre cómo describir e interpretar a tales robots y nuestras interacciones con ellos, y cómo deberíamos tratarlos y evaluar nuestras interacciones con ellos. ¿Qué está pasando exactamente cuando interactuamos con tales robots? ¿Cómo deberíamos responderles moralmente hablando? ¿Cómo deberíamos

comportarnos frente a ellos? ¿Cómo deberíamos percibirlos? ¿Qué sentimientos deberíamos albergar hacia ellos? ¿Cómo deberíamos hablarles? Por ejemplo, ¿es apariencia o realidad lo que está ocurriendo en estas interacciones? ¿Está bien sentir algo parecido a la empatía por los robots? ¿O es una equivocación, ya que los robots son «solo máquinas»? ¿Es mejor no apagarlos? ¿Por qué no? ¿Es aceptable ser «cruel» con ellos? ¿Qué es exactamente lo que tiene esto de malo? ¿Cómo deberíamos hablar con las máquinas? ¿Deberíamos ser amables con ellas, o podemos insultarlas todo lo que queramos? ¿Qué tipo de lenguaje y comportamiento podemos esperar de *ellas*?

Como veremos en este capítulo, estas cuestiones son relevantes no solo para reflexionar sobre robots en la ciencia ficción o en un futuro lejano. Hasta cierto punto, estas cuestiones ya nos plantean un desafío hoy en día.

¿ES ACEPTABLE AMAR O TORTURAR A LAS MÁQUINAS ?

Las máquinas que se parecen a los humanos y crean la apariencia de estar vivas tienen una larga historia cultural. Ya en la época antigua tenemos en las *Metamorfosis* de Ovidio la historia de Pigmalión, un escultor que se enamora de una estatua que él mismo ha esculpido. Desea que estuviera viva, y Venus le otorga su deseo: los labios de la escultura se sienten cálidos, y el marfil se vuelve suave. Pigmalión se casa con la escultura, que se transforma en una mujer. En el siglo XIX, tenemos el relato corto de E. T. A. Hoffmann «El hombre de arena» (Der Sandmann), en la que un hombre (Nathanael) se enamora de una mujer (Olimpia) que resulta ser un autómata. Aparte de las descripciones problemáticas relativas al género de estas historias y su relación con los robots sexuales contemporáneos (ver el capítulo 3), para los propósitos de este capítulo lo que es importante notar es que, en ambos casos, un objeto artificial, no vivo, parece estar vivo y ser humano. En la percepción y en la imaginación de las personas en este relato, se cruzan las fronteras de lo n o vivo/vivo y no humano/humano. Esto también ocurre en otro relato famoso de principios del siglo XIX, *Frankenstein*, aunque en este caso no es una máquina lo que vuelve a la vida, sino un cadáver. Sin embargo, en todos los casos los humanos usan ciencia, tecnología y oficio para crear algo que parece estar vivo y es similar a un humano. Deja de ser «solo una máquina» o un objeto inerte. Este cruce de fronteras atrapa a la imaginación, y suscita miedo y fascinación. Hay alguna duda sobre el estatus de estas entidades. ¿Están vivas o no? ¿Son solo máquinas o son algo más que máquinas? A

veces existe un sentimiento de extrañeza cuando nos encontramos con tales entidades, incluso si hay familiaridad con la forma y el comportamiento similar al humano. Usando el término de Freud de su ensayo de 1919 *Das Unheimliche*, en donde comenta sobre muñecos, podríamos decir que estas máquinas y criaturas son *inquietantes*. Diré más sobre esto en la sección siguiente.

Los robots que parecen y se comportan como humanos no son mera ciencia ficción. A pesar de que los robots actuales no son realmente capaces de engañar a las personas de modo que piensen que están interactuando con un ser humano, los llamados androides (robots diseñados con la finalidad de que parezcan y se comporten como humanos) ya se están desarrollando. Pero todos los tipos de robots, con forma humana (humanoide) o no, están creados para parecer seres vivos, y para parecer humanos o animales. Considérense los robots perros o los robots sexuales. De forma interesante, a pesar de saber y reconocer que estos robots son máquinas, las personas responden ante tales robots durante interacciones reales de formas que son sorprendentemente similares a como responderían a los seres humanos o a animales (no humanos), dependiendo de su apariencia y comportamiento. Por ejemplo, se dice que algunas personas son crueles con los robots, o que otras empatizan con ellos. Incluso a pesar de que la gente sabe que los robots son meras máquinas, interactúan con ellos como si fueran seres humanos, animales, criaturas fantásticas que tienen alguna similaridad con los humanos o animales, y así sucesivamente.

Analizaré este fenómeno ofreciendo cuatro casos/ejemplos concretos que sacan a la luz la dimensión ética:

1. ¿Es éticamente aceptable pegar una patada a un robot? La pregunta surgió en febrero de 2015 cuando la empresa robótica Boston Dynamics publicó un vídeo en el que los empleados le pegan patadas a un robot perro para mostrar cómo es de robusto. Un artículo de la CNN observó que «como los robots comienzan a actuar y a parecerse cada vez más a cosas vivas, es cada vez más difícil no percibirlos de esa forma. Mientras que, en principio, patear a un robot no supone abusar de una cosa viva, después de ver el vídeo muchas personas se sintieron inicómodas[1]». Sus comentarios en Twitter incluían cosas como «pobre Spot», «Dar patadas a un perro, incluso a un robot perro, simplemente parece estar mal», y «en serio, Boston Dynamics, ¿podéis dejar de dar patadas a esos pobres robots? ¿Qué os han hecho!?». ¿Es

problemática esta tendencia a antropomorfizar (atribuir cualidades humanas a entidades no humanas) y esta preocupación ética por robots «sin vida»? ¿Por qué, o por qué no?

2. A continuación sucedió la «decapitación» y «asesinato» de Hitch-BOT. HitchBOT fue diseñado como robot social con personalidad y búsqueda de personas que lo llevaran en autostop. «El» preguntaría a la gente si podrían llevar*lo* en el coche. Durante el verano de 2014, el robot hizo autostop a través de Canadá en un viaje de más de diez mil kilómetros. Cuando en 2015 una nueva versión hizo autostop en los Estados Unidos, su viaje acabó enseguida en Filadelfia, donde fue destrozado. Un artículo informó del «desbordamiento de sentimientos y de personas expresándolos a miles»[2]. «¿Puedes matar a un robot?», preguntaba un artículo de la BBC[3].

3. El trabajo psicológico muestra que la gente es capaz de empatizar con robots humanoides. Yutaka Suzuki y sus colegas compararon las respuestas neurales al «dolor» percibido del robot con las respuestas al dolor humano. Los investigadores mostraron imágenes de situaciones dolorosas y no dolorosas que implicaban manos humanas y robóticas, como un dedo que va a ser cortado por un cuchillo. A pesar de que las personas empanzaron más con los humanos que con los robots, en ambos casos hubo lo que los científicos interpretaron como respuesta empática: el dolor percibido parece suficiente para que empaticemos con robots humanoides[4]. Kate Darling también descubrió en sus estudios que la gente dudaba a la hora de «torturar» o aplastar robots. Por ejemplo, cuando se les pidió a los participantes en un experimento que golpearan al juguete robótico Hexbug Nano, dudaron en hacerlo cuando el robot se «presentaba dentro de un marco antropomórfico»: esto es, cuando se les contaba una historia de fondo que atribuía cualidades al robot que lo hacían parecer vivo, o cuando se le daba un nombre personal[5].

4. Un estudio realizado en Alemania investigó la reacción de las personas cuando se les daba la opción de apagar un robot que expresaba objeciones emocionales a que se lo apagara. El robot decía, «¡No! ¡Por favor, no me apagues! ¡Tengo miedo de que no me vuelvan a encender!». Después de esa protesta, aquellos que finalmente decidían apagarlo dudaban antes de hacerlo y experimentaron estrés posteriormente. Los investigadores concluyeron, «provocada por la objeción, las personas tendieron a tratar al robot como una persona real más bien que simplemente como una máquina, cumpliendo o, al

menos, tomando en consideración, si cumplir su petición de continuar encendido[6]».

¿Cómo deberíamos responder a estos fenómenos desde la posición de la ética de los robots? Este capítulo trata sobre la cuestión de la «paciencia moral»: ¿qué les debemos a los robots, si es que les debemos algo? Pero déjenme primero continuar discutiendo un tipo de robot que puede parecer que sugiere esta cuestión más que otros, y que plantea cuestiones sobre el significado moral de la apariencia, la ilusión y (de nuevo) el engaño: los androides.

¿DEBERÍA IMPORTAR ÉTICAMENTE LA APARIENCIA DE LOS ANDROIDES?

Los androides son robots que se parecen y comportan como humanos. Los robotistas como Hiroshi Ishiguro intentan crear robots de este tipo.

Actualmente la mayoría de los androides son marionetas controladas de forma remota con una apariencia humana y, al examinarlos más de cerca, son fácilmente distinguibles de los humanos. No superan el denominado test de Turing, que exigiría que fueran indistinguibles (al menos por un largo tiempo). Sin embargo, los androides con una apariencia muy humana pueden parecer humanos *a primera vista*. En un experimento se mostró a unas personas un androide durante dos segundos: el 70 % de los sujetos no se dio cuenta de que estaban tratando un con un androide, dado que el robot realizó micromovimientos[7]. Además, si tales robots estuvieran combinados con IA, incluyendo, por ejemplo, tecnología de procesamiento de lenguaje natural y, por tanto, siendo más autónomos e interactivos, parecerían más humanos y las personas tendrían probablemente reacciones emocionales (más intensas).



Figura 4. Ejemplo de un robot androide, ERICA. ERATO
Proyecto Ishiguro de interacción simbiótica humano-robot.

Esto puede deberse a la empatía, pero también al miedo o a una sensación extraña. Algunos robots son espeluznantes, especialmente los de la ciencia ficción. Si los androides fueran más semejantes a los humanos en apariencia y en comportamiento (en lugar de meramente humanoides), también plantearían ese problema: las personas tendrían una experiencia espeluznante e inquietante al encontrarse con tales entidades. El concepto de lo inquietante, usado ya por Freud, lo han tomado los propios robotistas: históricamente por el diseñador de robots Masahiro Morí y contemporáneamente, por ejemplo, por Karl MacDorman e Ishiguro[8]. Morí presenta la hipótesis de que la familiaridad se incrementa con el parecido humano, pero que, en un determinado punto, se encuentra lo que denomina un valle inquietante: los robots parecen humanos, pero no del todo. Hay una desviación sutil de la apariencia humana. Esto crea un sentimiento de extrañeza o inquietud. Por ejemplo, un animal disecado o un robot humanoide todavía no son inquietantes, ya que parecen algo familiar, a pesar de que no están cerca de parecerse a un humano. No se perciben como algo extraño; al contrario, algunas personas los quieren abrazar o hablar con ellos. Un cadáver, un zombi o, ciertamente, un androide futuro, sin embargo, tienden a provocar en la gente el sentimiento mezclado de familiaridad y extrañeza que Freud identifica con lo inquietante.

¿Cómo deberían evaluarse éticamente estos fenómenos? ¿Es éticamente aceptable, o bueno, crear robots con una apariencia humana, dado que pueden inducir a las personas, a pesar de que son seres humanos, y en el caso de los

androides, provocar en ellas el sentimiento de lo inquietante? ¿Si hubiera entre nosotros robots extremadamente parecidos a los humanos, cómo deberíamos comportarnos con ellos? ¿Por ejemplo, qué, aparte los derechos de propiedad/posesión, debería evitar que nos portáramos de una forma moralmente mala con estos robots? ¿Qué sería una forma «mala»? Si hay una forma mala, ¿cuál sería una «buena»?

He aquí algunas respuestas que pueden darse, y se han dado, a estas dos cuestiones.

En primer lugar, como hemos visto, varios autores han planteado el problema del engaño. Se podría defender que, en el caso de robots extremadamente parecidos a los humanos, este problema se vuelve incluso más urgente. ¿Si es realmente *tan* problemático inducir engañosamente a que las personas piensen que esos robots son humanos, como mantienen tanto los Sparrow como los Sharkey, entonces es éticamente aceptable apoyar el desarrollo de tales robots? Para evitar problemas con respecto al engaño y al valle de lo inquietante, se podría decidir no desarrollar robots extremadamente parecidos a los humanos, como los androides, o incluso los robots humanoides en general. O se podría, al menos, prohibir el uso de tales robots para grupos vulnerables específicos como los niños pequeños y algunas categorías de ancianos. Otra respuesta podría ser defender que el desarrollo de estos robots debería permitirse, pero que los diseñadores deberían ser francos sobre lo que es el robot y lo que puede hacer. Bruce Tognazzini, un diseñador de *software*, ya ha propuesto una ética de la honestidad con respecto al diseño de interfaces humano-ordenador; esto también se podría aplicar a los robots[9]. De acuerdo con esta posición, los diseñadores de robots deberían crear la tecnología de una forma en que los usuarios se volvieran conscientes de que el robot es solo una máquina. Si crean la ilusión de que es más que una máquina, deberían llamar a lo que hacen una actuación o una representación (*performance*). El diseño de los robots es, entonces, una especie de magia escénica, y la magia escénica tiene su propia ética. Los magos éticos no pretenden (al menos fuera del contexto de sus representaciones) que lo que hacen sea real. Su trabajo es crear ilusiones, y la gente es consciente de ello. Los diseñadores de robots deberían adoptar una ética similar.

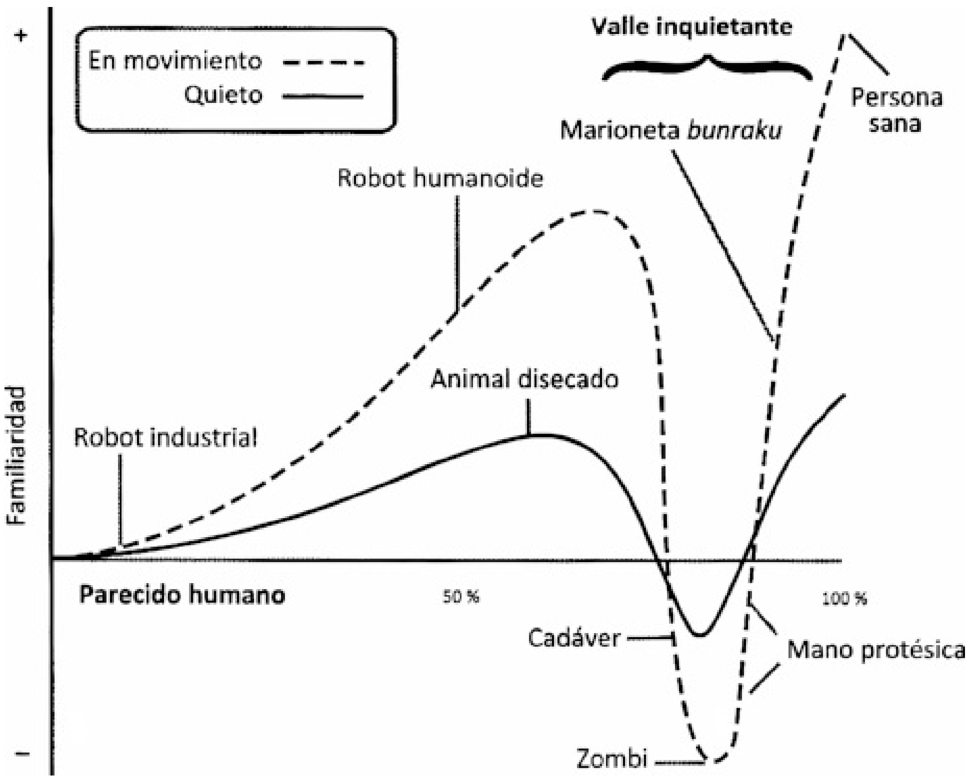


Figura 5. El valle inquietante. Fuente: Wikimedia Commons, 20 de noviembre de 2021, en: <https://commons.wikimedia.org/wikiFile: Mori_Uncanny_Valley _de.svg.>. Véase también Mori, «The Uncanny Valley», y MacDorman e Ishiguro, «The Uncanny Advantage».

Además, si el diseño y las máquinas están tan implicados en el engaño y en crear ilusiones, como ha señalado el filósofo del diseño Vilém Flusser, entonces se podría ir más allá y explorar qué significaría conceptualizar la ética de los robots como ética de la representación (*performance*)[10]. He sugerido recientemente la posibilidad de una ética de la representación para evaluar las tecnologías de la información y de las comunicaciones, particularmente la que no está necesariamente basada en la distinción entre la realidad y la ilusión, sino en otros criterios (¿qué es una buena representación?) y, tomando en consideración que otros usuarios de las tecnologías de la información y de las comunicaciones funcionan como co-magos[11]. Los usuarios no son meras «víctimas» de los diseñadores de robots supuestamente «malvados»; la ilusión es cocreada y, por lo tanto, los usuarios son también responsables de los efectos.

> El diseño de los robots es, entonces, una especie de magia escénica, y la magia escénica tiene su propia ética.
> Los magos éticos no pretenden que lo que hacen sea real. Su trabajo es crear ilusiones.

En segundo lugar, existen varias respuestas posibles a la cuestión de qué debemos a los robots, que también está relacionada con la de si y cómo se debería (des)contar la apariencia, así como la experiencia, en la ética de los robots. ¿Es suficiente decir que los robots son «meras máquinas», o existe una forma de tomarse seriamente y, al menos, *entender* mejor qué está ocurriendo en estas interacciones humano-robot? ¿Cuáles pueden ser las implicaciones normativas si hacemos esto? Discutiré con mayor profundidad estas cuestiones en términos del problema con respecto a la paciencia moral de los robots.

¿CUÁL ES LA POSICIÓN MORAL DE LOS ROBOTS?

Como ilustraron los ejemplos anteriores de este capítulo, las personas no siempre tratan a los robots como «meras cosas» o «meras máquinas». A veces con respecto a algunos robots se preocupan, empatizan, temen, tienen un sentimiento extraño, etc. Considérese de nuevo los casos en los que se daban patadas a un robot o en los que se empatizaba con ellos. ¿Deberían estas experiencias, interacciones y comportamientos descartarse o no tenerse en cuenta moralmente (por ejemplo, ¿está mal sentirse de esta manera o tratar al robot de aquella otra, o es una experiencia que puede ignorar la filosofía moral?), o hay algo correcto en estas respuestas y los filósofos morales deberían prestarles atención? ¿Cómo debería responder a estos fenómenos la ética de los robots que plantea preguntas *normativas*? Esta cuestión puede abordarse preguntándose no si los robots pueden ser *agentes* morales, como

en el capítulo anterior, sino más bien, si pueden ser *pacientes* morales. ¿Qué es lo que se les debe *a* los robots, si es que se les debe algo? O, en otras palabras, ¿cuál es la *position moral* de los robots?

La cuestión con respecto a la posición moral directa

La manera en que se formula ahora esta pregunta concierne a lo que denominamos posición moral «directa». La cuestión es entonces si existen algunas propiedades intrínsecas del robot que justifiquen otorgarle una posición moral en particular. ¿Existe algo en y sobre el robot que pueda justificar otorgarle una posición moral? ¿O es simplemente una cosa y, por lo tanto, deberíamos tratarlo como tal, independientemente (y a pesar) de las ilusiones que puedan querer crear los diseñadores cual magos escénicos?

La posición más común y, quizás, más intuitiva es que los robots son cosas, ya que (en contraste con muchos animales) carecen de propiedades como la conciencia, la sensibilidad y no tienen estados mentales. Joanna Bryson cree que los robots son herramientas y propiedad nuestra y, así, no tenemos obligaciones con ellos: deberían ser nuestros siervos[12]. Aquellos que defienden esta posición puede que también estén de acuerdo con que, *si* los robots fueran conscientes, tuvieran estados mentales, etc., deberíamos otorgarles una posición moral, pero no creen que los robots actuales cumplan estos criterios.

Sin embargo, un problema potencial de esta posición es que rio explica o justifica las experiencias morales como las que se han relatado antes en este capítulo. Defiende que las personas están equivocadas pensando que tienen esas experiencias. Se las considera ilusorias, no reales. Esta posición también tiende a desestimar cualquier discusión sobre la posición moral del robot como irrelevante ante preocupaciones más urgentes; se la descarta como perteneciente solo a la ciencia ficción. Por el contrario, David Gunkel y yo mismo hemos afirmado sistemáticamente que tiene sentido y es filosóficamente interesante, como mínimo, *plantear* (y criticar) la cuestión con respecto a la posición moral de los robots. Por ejemplo, inspirado por Emmanuel Levinas, Gunkel ha planteado si se les puede otorgar derechos a los robots, o ser considerados como un «otro» que merezca dignidad: esto es, apelar a nuestra responsabilidad ética[13]. Sin embargo, pocas personas defienden directamente los derechos de los robots o posiciones similares que les otorguen una posición moral directa (una excepción es John Danaher; ver más abajo). ¿Cambiará esto en el futuro cuando los robots puedan mejorar la

creación de apariencias de propiedades humanas? Algunos creen incluso que en el futuro los robots *tendrán* propiedades como la conciencia. Por ahora, sin embargo, la posición convencional de que los robots no tienen una posición moral mayor que una tostadora o un lavavajillas prevalece y parece plausible. De acuerdo con esta posición, patear a un robot es como patear a otras máquinas que usamos, como coches y lavadoras: *puede* que esté mal si lo que pateamos es propiedad de otra persona, pero, aparte de eso, no hay ninguna razón para no hacerlo, o al menos ninguna razón que tenga que ver con las propiedades intrínsecas del robot.

LA CUESTIÓN CON RESPECTO A LA POSICIÓN MORAL INDIRECTA

Si no existen buenas razones para otorgar a los robots una posición moral, dado que se cree que son meras cosas, ¿no existe una forma de justificar *otras* intuiciones y experiencias morales, como el sentimiento de que hay algo erróneo en patear a un robot? Hay una vía interesante para este propósito: no se trataría de defender una posición moral directa, sino una *indirecta*. Aquí, la idea es que debemos tratar al robot de forma correcta no porque tenga ciertas propiedades particulares, sino porque nosotros los *humanos* tenemos una posición moral y, por tanto, necesitamos ser tratados de forma correcta. Considérese cómo argumentaba Immanuel Kant en relación con los perros. De acuerdo con Kant, disparar a un perro está mal, no porque viole ninguna obligación hacia el *perro*, sino porque una persona que realiza este acto «daña las cualidades amables y humanas de él mismo, que debería ejercer en virtud de sus obligaciones con la humanidad[14]». Lo que cuenta en este razonamiento no es el robot: son el carácter y el bien de los humanos y, en consecuencia, *sus* propiedades morales relevantes. De acuerdo con Darling, este razonamiento «se extiende al trato que damos a los compañeros robóticos[15]». Se podría aplicar al trato que dispensamos a todos los robots. El problema de pegar patadas a un robot o de hacerles cosas «malas» no surge porque les debamos nada, sino porque debemos tratar a las personas de forma correcta. La preocupación es que, si alguien «maltrata» a un robot, esto puede conducir al maltrato a humanos (un argumento consecuencialista) y/o a que los humanos no ejerzan sus obligaciones hacia otros humanos (un argumento deontológico) y/o a un carácter moral malo y vicioso (un argumento de la ética de la virtud) de parte de los humanos (la formulación exacta o el énfasis depende de la teoría normativa moral de la cual partamos). En otras palabras, patear a un robot es malo no porque se cause ningún daño al robot, sino

porque tal comportamiento puede extenderse hasta el trato con los humanos, y/o porque viola las obligaciones que tenemos hacia otros humanos, y/o porque fomenta un carácter moral malo de parte de ellos: no es virtuoso.

> Patear a un robot es malo no porque se cause ningún daño al robot, sino porque… fomenta un carácter moral malo de parte de ellos: no es virtuoso.

La ventaja de este argumento en favor de la posición moral indirecta es que explica y justifica la intuición de que haya algo malo al maltratar a los robots, especialmente a los robots con apariencia humana y los sociales. Explica qué es exactamente lo que está mal, y nos ofrece una base para evaluar el comportamiento humano y los sentimientos hacia los robots. Si tomamos esta posición, no es simplemente ilusorio, moralmente engañoso o irrelevante sentir que hay algo de malo cuando alguien tortura a un robot. Está bien, e incluso está *justificado*, sentirse así, ya que (usando un argumento sobre la posición moral indirecta en combinación con el consecuencialismo, la deontología y/o la ética de la virtud) disponemos ahora de un buen argumento (o para los pluralistas morales, varios) para condenar moralmente, o al menos cuestionar, tal comportamiento.

PENSAR DE OTRA MANERA Y EL ENFOQUE RELACIONAL DE LA POSICIÓN MORAL

Sin embargo, este no es el fin de la discusión. Las cosas son más complejas y problemáticas de lo que parecen a primera vista. Para empezar, se podría criticar la actitud y los procedimientos mismos que se usan en estos argumentos a favor o en contra de la posición moral directa e indirecta.

Por ejemplo, nuestro razonamiento sobre la posición moral se basa típicamente en las premisas que dicen que una entidad tiene propiedades

intrínsecas particulares (y que las propiedades particulares tienen un significado moral particular). Normalmente tiene la siguiente forma:

> Solo las entidades del tipo A con propiedades P, Q, R… tienen una posición moral S.
> La entidad X es del tipo A (y tiene las propiedades P, Q, R…).
> Por lo tanto, la entidad X tiene la posición moral S.

Por ejemplo:

Solo las entidades conscientes tienen derechos morales.
Este robot no es consciente.
Por lo tanto, este robot no tiene derechos morales.

Para empezar, puede haber otras propiedades distintas de la consciencia que justifiquen los derechos morales. Pero si cambiamos «solo» por «todas» en la primera premisa, entonces el argumento no funciona, ya que la conclusión no se sigue. Ahora bien, si arreglamos esto y cambiamos la primera premisa por «la consciencia es una condición suficiente para tener derechos morales», entonces debemos continuar cuestionando las premisas y preguntar: ¿cómo podemos tener por cierto que la consciencia es una condición suficiente para tener derechos morales (la primera nueva premisa), y cómo tenemos por cierto que los robots, o cualquier otra entidad, tienen o no propiedades relevantes, tales como la consciencia (la segunda nueva premisa)? En la práctica, tendemos a decidir esto en base a las apariencias, incluso en el caso de los humanos. Damos por sentado simplemente que los humanos tienen consciencia basándonos en las apariencias y en el comportamiento. Sin embargo, como filósofos, podemos y debemos plantear preguntas escépticas sobre cuál es la base de estas suposiciones. Nótese también que los filósofos tienden a estar en desacuerdo sobre lo que es la consciencia. ¿Cómo tenemos por seguro que una propiedad particular garantiza un estatus particular? ¿Cuál es el fundamento de este conocimiento? ¿Se ha decidido pura y simplemente esto por nosotros (digamos, una cultura, sociedad o comunidad particular), o existen razones más sólidas?

Es más, parece haber algo erróneo en la actitud y el procedimiento que usamos cuando, como filósofos morales o responsables de la toma de decisiones, participamos en un razonamiento moral sobre el estatus de otras

entidades. ¿Es correcto que nos coloquemos (como filósofos, políticos o humanidad) en la posición de juez supremo de la posición moral de otras entidades? Realizamos una especie de disección moral de esa entidad como monarcas del reino moral. ¿Qué justifica que adoptemos esta posición jerárquica y dominante? ¿Adoptar esta posición no decide ya de antemano algo sobre la posición moral de otras entidades, incluso *antes* de que comencemos nuestro razonamiento moral?

Además, parece que la posición moral que otorgamos a otras entidades depende de condiciones subjetivas y sociales, está influenciada por el lenguaje que usamos, y varía histórica y culturalmente. Por ejemplo, ahora mucha gente cree que (algunos) los animales tienen derechos, pero este no foe siempre el caso. Algunos animales son tratados de forma diferente si se les da un nombre propio y el estatus de «mascota», mientras que otros (a veces el mismo animal biológico) son maltratados y sacrificados en otros contextos. Comparemos un cerdo que se tiene como mascota y un cerdo que se come. Mientras que algunas personas comen perros, otras los tratan como miembros de la familia. El cómo tratamos a los robots parece también corresponderse con (si es que no se basa sobre ello) cómo hablamos sobre y con el robot[16]. Una vez que le damos un nombre, tendemos a tratarlo de una forma diferente. ¿Qué implica o debería implicar todo esto para nuestra reflexión sobre la posición moral?

Haciéndome precisamente estas preguntas, he propuesto un enfoque «relacional» de la posición moral y he investigado sobre las condiciones de posibilidad de la adscripción de una posición moral[17]. Dado el escepticismo sobre lo que he llamado «el enfoque de las propiedades» y la distancia creada por el procedimiento moral de decidir jerárquicamente sobre la posición de otras entidades, he planteado si podemos reflexionar de una forma más relacional y crítica sobre los robots, siendo más cautelosos cuando pensamos que conocemos la posición de otras entidades, y siendo conscientes de que el lenguaje y las relaciones que tenemos realmente con estas entidades conforman cómo pensamos sobre ellas. Para el razonamiento crítico sobre la posición moral de los robots, esto podría implicar que nos tomemos en serio y reflexionemos adicionalmente sobre el fenómeno de que lo que creemos que «es» un robot y que cómo lo tratemos depende de cómo esté construido y del establecimiento de una «relación» con ese robot, incluyendo el lenguaje que usamos con él, y que todo esto influye sobre lo que pensamos sobre su posición moral. Considérense de nuevo los experimentos de Darling, que

pueden ser interpretados como una muestra de cómo una narrativa y un lenguaje particulares construyen al robot «personal» y, por lo tanto, tienen implicaciones sobre cómo las personas lo tratan.

A pesar de que estas observaciones no nos dicen nada normativo de forma directa sobre la posición moral del robot (es una cuestión descriptivo-hermenéutica frente a otra normativa: describe y nos ayuda a entender los fenómenos, pero no nos ofrece una guía directa sobre lo que hacer), nos debería hacer, al menos, que fuéramos más cautelosos cuando adscribimos una posición moral. Si es así como funciona la adscripción de una posición moral, entonces quizás debamos ser *menos* directos e impulsivos cuando realizamos afirmaciones sobre la posición moral de otras entidades como los robots. Al final, adoptar un enfoque relational implica que debemos plantearnos la propia pregunta de la posición moral. He defendido que deberíamos considerar cómo se adscribe la posición moral y qué hace posible esta adscripción[18]. Gunkel, simpatizando con el enfoque relational, ha defendido que necesitamos una forma de pensar más orientada hacia el otro que no aborde la cuestión desde el punto del yo, sino desde el del (el de la dignidad del) otro[19]. La cuestión entonces no es qué «es» el otro, sino cómo puedo responder al otro. Preguntarse si los robots pueden ser otros desde un enfoque de las «propiedades» puede ya ser una «violencia» para el otro.

Lo que implica exactamente este enfoque «relational» y crítico para la ética de los robots normativa aún está en discusión[20]. He defendido que, en cualquier caso, es conveniente más precaución, paciencia y análisis crítico en cuestiones que conciernen al estatus moral de entidades en general[21]. Con respecto a la cuestión del engaño, he propuesto que la forma en que realmente usamos y respondemos a los robots puede ser descrita usando la metáfora de la representación[22]. Se podría decir que una situación similar a la del test de Turing sustituye los criterios de realidad con criterios de representación: se convierte en una especie de magia escénica de lo que podríamos llamar, siguiendo a Walther Zimmerli, una búsqueda «cartesiana» moderna para distinguir entre la realidad y la ilusión, para evitar un engaño maligno[23]. Sin embargo, no está claro qué implique este giro representativo y, quizás, también nietzscheano, para el estatus moral de los robots, normativamente hablando. ¿Qué implicará exactamente una ética de la representación (robótica)? Este punto se merece más investigación. Danaher ha propuesto, por el contrario, una teoría normativa que defiende directamente otorgar a los

robots un estatus moral si son «aproximadamente equivalentes desde el punto de vista representativo» a otras entidades que tengan un estatus moral significativo[24]. Lo que ocurre «dentro» no importa para el estatus moral; lo que importa es si el robot pasa un umbral representativo. Danaher defiende, así, una forma de conductismo ético.

La discusión sobre el enfoque relational también invita a que nos planteemos la cuestión de cuáles son las implicaciones normativas, si es que las hay, para las observaciones de que nuestras ideas sobre la posición moral tienden a cambiar, como lo hacen nuestras ideas sobre los robots humanoides. Con respecto al cambio de ideas sobre la *posición moral*, puede ser interesante contemplar la historia de nuestras ideas sobre la posición moral de los animales. Hoy en día a muchos animales se les da una posición moral superior de la que recibieron en el pasado. Y, por supuesto, las discusiones filosóficas sobre el relativismo moral son aquí relevantes. Con respecto al cambio de ideas sobre los *robots humanoides*, Julie Carpenter coloca nuestras respuestas psicológicas a los robots en una perspectiva sociocultural e histórica[25]. Yendo más allá de Mori, ella propone lo que llama «la teoría del proceso de acomodación del robot» para comprender cómo respondemos a los robots como sociedad. Al principio, los robots eran un concepto nuevo, se veían como entidades inquietantes, y se crearon nuevas historias y mitos. Pero, a continuación, después de que viéramos a los robots en la ciencia ficción y de que efectivamente se produjeran en masa, los robots se han vuelto más comunes, integrados en un sistema social, y han surgido normas con respecto a ellos. En línea con Carpenter, se podría decir que, durante este proceso, los robots se vuelven *normales* y *normalizados*. Con el tiempo, dejaremos de temerlos y, en su lugar, los aceptaremos. Algunas personas incluso quieren tener sexo con ellos. Y, hoy en día, a medida que continuamos negociando cómo vivir con robots y cómo integrarlos más aún en nuestra sociedad y cultura, quizás también cambiemos nuestras opiniones sobre su posición moral, por ejemplo, otorgándoles una posición moral más elevada que a las cosas y desarrollando nuevas normas sobre cómo tienen que ser tratados, o considerarlos como «cosas» normales entre otras cosas. Se podría decir, por tanto, que nuestras *relaciones* con el robot es probable que cambien con el tiempo.

Sin embargo, sea cual sea el resultado, continúa siendo dudoso qué implica reconocer este cambio histórico, social y cultural a la hora de preguntarse, así como de responder, a la cuestión normativa relacionada con

la posición moral. Al menos se pueden aprender algunas lecciones generales. Dado el cambio histórico y la elaboración del enfoque relacional, repito mi recomendación de que, al menos, deberíamos ser cautelosos cuando adscribimos una posición moral a los no humanos. Quizás deberíamos estar menos seguros sobre que nosotros mismos, como humanos, conozcamos las respuestas. Y quizás debiéramos ser menos arrogantes sobre nuestras *preguntas*. Después de todo, al plantear la cuestión con respecto a la posición moral de los no humanos, asumimos, de hecho, una posición jerárquica y distante desde la que juzgamos a nuestros «subalternos», que se encuentran lejos y más abajo de nosotros. Pues la forma en que planteamos la pregunta ya categoriza a los que juzgamos. Sin cautela y modestia y, ciertamente, sin cambios en las preguntas que planteamos, la filosofía moral se arriesga a perpetuar una historia de dominación, violencia y colonización con respecto a los no humanos. Superar este problema es un desafío no solo para la ética medioambiental y animal, sino también para la ética de los robots.

Al plantear la cuestión con respecto a la posición moral de los no humanos, asumimos, de hecho, una posición jerárquica y distante. Pues la forma en que planteamos la pregunta ya categoriza a los que juzgamos.

Para concluir: la cuestión que concierne a la posición moral de los robots ha conducido y está conectada con las respuestas críticas al proyecto filosófico general de la adscripción de una posición moral, y ha desafiado no solo a respuestas dogmáticas a la cuestión, sino a la cuestión misma. Mirando de una forma menos profunda hacia el abismo abierto por este ejercicio de escepticismo e investigación filosófica crítica, recomiendo una mayor

discusión sobre las implicaciones normativas más prácticas de un enfoque relacional, y cómo las diferentes teorías normativas (por ejemplo, la ética de la virtud: véase mi obra reciente) y las teorías metaéticas influyen en esta cuestión (a pesar de que la cuestión del relativismo nos puede conducir de nuevo al abismo)[26]. Tratando también y siempre sobre los humanos y su ética, la ética de los robots continúa, y debe continuar, en diálogo con una ética y filosofía moral general, desde la que pueda aprender y a la que pueda contribuir.

Capítulo 7

Drones asesinos, distancia y la existencia humana

> Estás viendo a gente moviéndose alrededor, gente arrastrándose, y entonces, se produce una gran explosión: «algunos están muertos, otros corriendo y algunos reptando, todo es bastante salvaje», recordaba. «Muchos de ellos eran helicópteros, y los podías oír, con lo que a veces, [las personas que estaban en el suelo] evitando ser vistos trataban de esconderse en los árboles, eludirlos dando rodeos entre los edificios, y cosas por el estilo».
>
> Justin Rohrlich, «Why a U.S. Soldier»

Estas palabras son de Essam Attia, un exsoldado involucrado en la guerra de drones norteamericana en Irak que ahora hace campaña contra los «drones asesinos». El término suena como si fuera ciencia ficción y nos puede hacer recordar películas como *Terminator*, pero los llamados «vehículos aéreos no tripulados» [*Unmanned aerial vehicle*, UAV] con capacidades letales no son robots humanoides, sino aeroplanos sin piloto humano a bordo que se pueden utilizar para distintos fines, incluyendo la vigilancia y el asesinato[1].



Figura 6. Dron Asesino.

Los ha usado el Ejército de los Estados Unidos al menos desde 2001, cuando la CIA intentó asesinar a un líder talibán y comenzaron los ataques de drones en Afganistán. Desde entonces, miles de personas, entre ellas civiles, han sido asesinadas por ataques de drones. Otras personas que han experimentado los ataques en tierra sufren angustia mental y problemas de corazón, a menudo sin recibir tratamiento para enfermedades mentales. Consecuentemente, ha habido campañas contra el uso de los llamados drones asesinos. No solo los Estados Unidos, sino también muchos otros países, usan, tienen o desarrollan estos drones.

Por ahora, los drones están siendo controlados de forma remota. Los pilotos de drones y otros operadores están sentados sobre el terreno en una oficina, por ejemplo, en los Estados Unidos, mientras que el «objetivo» puede encontrarse a miles de kilómetros, por ejemplo, en Afganistán. Los pilotos están mirando a las pantallas y tienen *joysticks* que parecen de videojuego. Pero no es un juego: se está matando a personas a distancia. Hoy en día, los operadores humanos aún deciden cuándo disparan los drones, pero se están desarrollando drones letales más autónomos, drones que no necesitan de un humano en la cadena. Con la tecnología de reconocimiento facial de IA y los sistemas de decisión se vuelve más sencillo construir sistemas que no solo puedan volar de forma autónoma (que ya es posible), sino también apuntar y matar de forma autónoma[2]. Los activistas y los académicos presionan para prohibir estas armas. Argumentan que hacen que comenzar una guerra sea más fácil, que no es justo que las máquinas decidan sobre las vidas humanas, y que la autonomía de las máquinas crea problemas para la atribución de responsabilidad. Por ejemplo, la campaña para detener robots asesinos (*Campaign to Stop Killer Robots*) intenta que los gobiernos y las organizaciones internacionales prohíban las armas completamente autónomas[3]. En respuesta, algunos países como Rusia y los Estados Unidos, que invierten en el desarrollo de estas armas, se han resistido a esta prohibición.

No solo los drones militares plantean cuestiones éticas. Los UAV se usan, por ejemplo, para vigilancia policial, y plantean por lo tanto cuestiones de privacidad y vigilancia. Los drones también se usan para acciones criminales, de terrorismo y de protesta. En diciembre de 2018, los drones causaron una gran perturbación en el aeropuerto de Gatwick, y en septiembre de 2019,

activistas medioambientales fueron arrestados por intentar alterar el tráfico aéreo en el aeropuerto de Heathrow[4]. Además, más allá de los drones, existen otros sistemas de armas letales autónomas que se usan en organizaciones militares que crean problemas éticos similares, como los sistemas de misiles autónomos. Este capítulo se centra en las armas letales autónomas militares, en particular, los UAV con capacidad letal (drones asesinos).

EL USO DE ROBOTS PARA FINES MILITARES

Históricamente, el desarrollo de tecnologías autónomas siempre ha tenido conexiones con el Ejército y la guerra. Considérese, por ejemplo, el desarrollo de la cibernética por Norbert Wiener, que recibió financiación del Gobierno de los Estados Unidos durante la Segunda Guerra Mundial para desarrollar armas antiaéreas capaces de apuntar y disparar: la finalidad era encontrar una forma de predecir la posición de los objetivos aéreos. O en cómo ayudó Alan Turing a las fuerzas aliadas en la Segunda Guerra Mundial descifrando el código de la Armada alemana. Los robots también se pueden considerar en el contexto de la historia del perfeccionamiento humano con fines militares, buscando «mejorar» a los guerreros humanos: usar robots autónomos es, entonces, un paso más, que busca eliminar por completo de la cadena al humano. También existen robots no autónomos bajo control remoto que se están usando por el Ejército, como los robots para la vigilancia o limpieza de minas.

> Históricamente, el desarrollo de tecnologías autónomas siempre ha tenido conexiones con el Ejército y la guerra.

Existen todo tipo de robots militares autónomos: robots terrestres tales como los vehículos blindados y los robots para eliminar minas, robots aéreos

que van desde pequeños drones hasta grandes aeronaves como el UAV con el escalofriante nombre de *Reaper* (segadora), usada por el Ejército de los Estados Unidos, robots marinos que incluyen vehículos subacuáticos no tripulados que pueden utilizarse en vigilancia y acciones submarinas, y vehículos espaciales autónomos. Existen sistemas autónomos que son fijos, pero, aun así, tienen partes móviles, como los sistemas antimisiles de los barcos: estos también pueden ser considerados un tipo de robot[5].

Los robots y los sistemas autónomos pueden tener varios grados de autonomía. En contextos técnicos militares, es normal realizar una distinción entre los humanos que están «en la cadena», que implica que los humanos tienen control directo y tienen algo que hacer antes de que pasen cosas (por ejemplo, presionar un botón para disparar), humanos «sobre la cadena» (humanos que supervisan y pueden intervenir vetando, por ejemplo, una decisión para atacar) y «fuera de la cadena». En esta última categoría la máquina puede tomar sus propias decisiones sin interacción con un humano: esta es la automatización completa. Actualmente casi todos los sistemas usados en las organizaciones militares están en o sobre la cadena. Pero se están desarrollando máquinas más autónomas, creando, por lo tanto, la opción de fuera de la cadena.

En la medida en que tales armas y robots tengan capacidades letales y, al mismo tiempo, se vuelvan cada vez más *autónomos* y más capaces de crear escenarios fuera de la cadena, se plantean varias cuestiones éticas. Si toman sus propias decisiones a la hora de apuntar y matar, ¿se les debería proporcionar a estos sistemas alguna capacidad para la toma de decisiones morales, o es esta una pregunta equivocada? Cuando los humanos ya no están en o sobre la cadena, ¿quién es responsable si algo sale mal, como cuando se mata a civiles? ¿O existen razones para que no queramos en absoluto disponer de esos sistemas? ¿Facilita el asesinato la guerra con drones? ¿Está mal matar a gente con máquinas, si es que se pueden justificar tales muertes? Las cuestiones concernientes a la agencia moral y a la responsabilidad se han discutido en el capítulo 5. El presente capítulo vuelve a estas cuestiones, pero también se pregunta sobre otros problemas éticos que tienen que ver con el uso de armas autónomas letales, en particular, con robots/drones asesinos, y discute si (no) deberíamos usarlos y por qué.

CUESTIONES ÉTICAS Y OTROS PROBLEMAS FILOSÓFICOS PLANTEADOS POR LOS DRONES ASESINOS

El argumento *a favor* del uso de UAV autónomos letales es que reduce las muertes de la parte atacante. Reducir las muertes es bueno en sí mismo, por supuesto, e incrementa el apoyo político para las operaciones militares. Como señala un artículo de *Foreign Affairs*, «ninguna bolsa para cadáveres vuelve a casa en virtud de una operación con drones[6]». Este razonamiento también se usa por los robotistas implicados en el desarrollo de tales armas. Ronald Arkin, por ejemplo, defiende el uso de los robots en la guerra diciendo que reduce las bajas de combatientes y no combatientes. Además de utilizar lo que podemos llamar el «argumento de la bolsa para cadáveres», afirma que tales sistemas pueden comportarse potencialmente de una forma *más* ética que los humanos en el campo de batalla, ya que las máquinas no se ven obstaculizadas por las emociones y pueden monitorizar de forma objetiva el comportamiento de todas las partes[7]. De acuerdo con esta perspectiva, necesitamos centrarnos en hacer que las máquinas sean más éticas. Considérese de nuevo el proyecto de crear «máquinas morales» mencionado en el capítulo anterior.

Contra esta perspectiva, sin embargo, se debería preguntar, «¿Menos muertes en qué lado?». ¿Qué pasa con las muertes del lado del «enemigo», y del sufrimiento de las personas cercanas que están conectadas con los que resultan asesinados y heridos, como la familia y los amigos y, ciertamente, incluso con el sufrimiento de quienes lo presencian junto con aquellos que tienen que vivir a diario bajo la *amenaza* y el riesgo de ataques con drones? Si se usa el argumento utilitarista, se debería al menos contar *esas* bolsas para cadáveres de *esas* personas. El argumento «no hay bolsas para cadáveres» solo funciona bajo la suposición de que está justificado comenzar la guerra particular en cuestión y usar drones en la guerra en esta forma concreta (asesinato selectivo, así como asesinato a distancia). Como veremos, esto se encuentra lejos de ser autoevidente. Además, según hemos visto en la discusión sobre la agencia y la responsabilidad morales, hay buenas razones en contra de la concepción de las máquinas morales ejemplificada por Arkin. Mostraré que también existen buenos argumentos contra el asesinato automatizado.

Otro argumento para usar armas autónomas letales en general es que en situaciones específicas puede ser necesario reaccionar rápidamente. Si los humanos son demasiado lentos para reaccionar a un ataque, entonces parece justificable usar sistemas autónomos para la autodefensa. Considérese el caso de los sistemas de defensa con misiles: si los humanos tienen que tomar

decisiones, puede que sea ya demasiado tarde. Contra este argumento se podría decir que tales razones conducen a una carrera para desarrollar las más eficientes armas autónomas. Este problema también nos recuerda el alcance global del problema: dado que el *software* se puede mover fácilmente a través de fronteras, y dado que los drones se pueden desarrollar y usar en cualquier parte, los problemas éticos tienen un alcance global. Consecuentemente, la prohibición o cualquier otra regulación solo tiene sentido si se hace con un acuerdo internacional. Esta es una razón por la que los activistas a menudo se centran en las Naciones Unidas.

Estos son dos argumentos a favor, pero existen muchos problemas éticos con los drones asesinos que apoyan argumentos *contra* su desarrollo y su uso. He aquí una visión general de los principales problemas.

LA GUERRA JUSTA Y LA RESPONSABILIDAD

En primer lugar, se ha cuestionado si las guerras en las que se usan los drones asesinos son «guerras justas». Si la guerra es moralmente aceptable, surgen entonces las cuestiones sobre si está justificada y cómo debería conducirse. La teoría de la guerra justa proporciona criterios con respecto a las condiciones para comenzar una guerra (*ius ad bellum*) y principios que deberían seguirse en el curso de las actividades bélicas (*ius in bello*). Con respecto al *ius ad bellum*, la preocupación está en que el uso de los drones asesinos puede hacer más sencillo comenzar una guerra. Los drones son más baratos y más fáciles de desplegar y, como ya se señaló, su uso tiende a encontrarse con menos resistencia pública debido a que desaparece el problema de las bolsas para cadáveres. Las luchas de drones son consideradas relativamente seguras y de poco riesgo[8]. Este hecho puede hacer que sea tentador para los políticos comenzar una guerra.

Cuando se trata del *ius in bello*, es un principio ético y legal generalmente aceptado que está mal matar a civiles no combatientes: está considerado un crimen de guerra. Pero los drones autónomos y el contexto en el que se usan (por ejemplo, en ciudad y en guerra de guerrillas) puede hacer que sea difícil distinguir entre civiles y combatientes. Arkin puede responder aquí que los drones autónomos pueden ser mejores a la hora de realizar esa distinción. Y Vincent Müller ha defendido que los robots reducen los crímenes de guerra, ya que no violan ni se enfadan, y siguen las órdenes al pie de la letra[9]. ¿Pueden ser los robots más éticos que los humanos en el campo de batalla,

como ha planteado John Sullins[10]? Esto se puede interpretar como una cuestión de principios sobre la capacidad de las máquinas para la inteligencia y la agencia moral (ver más abajo), pero si, en su lugar, usamos un razonamiento moral consecuencialista, también puede entenderse como una cuestión empírica que nos pide comparar el comportamiento humano con el de un robot en el campo de batalla (y sus consecuencias). Partiendo de este enfoque, se podría aceptar el punto de vista sobre los robots de que no son capaces de cometer crímenes de guerra específicos (por ejemplo, violar), pero discutir que, mientras que la tecnología no sea lo suficientemente buena como para realizar las distinciones necesarias, el principio *ius in bello* con respecto a discernir entre combatientes/no combatientes necesita que lo dejemos en manos de los humanos. Si son mejores que las máquinas, y mientras lo sean, los humanos deberán hacer la distinción: en caso contrario, acabaríamos teniendo crímenes de guerra. También se puede dudar de si el comportamiento no emocional es necesariamente y siempre más moral y responsable. Una concepción alternativa es que las emociones pueden asesorar al razonamiento moral y ayudar a ejercer la responsabilidad moral. Por ejemplo, la empatía hacia los civiles puede evitar crímenes de guerra. Además, junto con la responsabilidad del personal militar y de los políticos que toman las decisiones sobre ir a la guerra, deberíamos considerar la responsabilidad de los fabricantes de drones. Edmund Byrne ha mantenido que, dada la importancia de la inmunidad de los ciudadanos y que los fabricantes de los drones asesinos facilitan prácticas que matan a ciudadanos inocentes, las empresas deberían dejar de fabricar tales drones[11].

Además, Robert Sparrow ha defendido que el *ius in bello* necesita que la responsabilidad se pueda adscribir en casos en los que esté involucrado el uso de los robots asesinos y otros sistemas autónomos de guerra, especialmente cuando ese uso amenace vidas humanas. Pero ¿a quién debemos adscribir la responsabilidad? El *programador* y su empresa solo pueden ser responsables si existe una negligencia. Sin embargo, si dejan claras las limitaciones del sistema, entonces podrían salir del paso. Puede ser inherente a la autonomía del sistema que realice elecciones impredecibles que puedan no ser previstas. Puede hacerse responsable al *oficial al mando*, pero el problema es que se le hace responsable de acciones que no controla y puede ser castigado injustamente. Por diversas razones que tienen que ver con la falta de una agencia moral, no podemos mantener que las *máquinas* sean responsables (ni podemos castigarlas). En otras palabras, nos encontramos, de nuevo, con la

brecha de responsabilidad y el problema de la adscripción de responsabilidad. Parece que para los atacantes, no existe una forma de cumplir con su obligación para con los combatientes enemigos (y debemos añadir, no combatientes) de asegurar que alguien es responsable de lo que hacen los drones autónomos en el campo de batalla. Sparrow concluye que el empleo de tales sistemas es, por lo tanto, injusto[12]. Esta es una de las razones por la que muchos académicos y activistas han exigido la prohibición de los robots asesinos.

Aun así, no todo el mundo está de acuerdo con que la existencia de una brecha de responsabilidad sea suficiente para sustentar la conclusión de que no se deberían utilizar tales sistemas (por ejemplo, prohibirlos). Müller defiende que podemos intentar estrechar la brecha de responsabilidad mediante el uso de regulación y estándares. Deberíamos desarrollar estándares técnicos de fiabilidad, y hacer responsables a quienes cometen fallos de fabricación, distribución o despliegue de los sistemas de acuerdo con los estándares. Mantener una cadena de mando clara ayuda en este sentido. Usando el razonamiento utilitarista, Müller concluye que las consecuencias generales de los robots asesinos son positivas[13].

Esta es una afirmación controvertida, y pronto discutiremos más razones que nos harán dudar de ella (además de ofrecer otros argumentos deontológicos y de la virtud en contra). Pero incluso si está en lo correcto, asegurarnos de que se pueda adscribir la responsabilidad es una exigencia y regulación ética importante, y una cadena de mando clara puede no resolver el problema por completo. La cuestión es, entonces, qué principio podría servir de guía a la hora de lidiar con estos problemas de responsabilidad. En respuesta a la brecha de responsabilidad, las voces críticas en el debate legal y político han defendido un significativo control humano[14]. Sin embargo, existe cierta discusión sobre lo que esto implica, filosófica y prácticamente. Por ejemplo, Filippo Santoni de Sio y Jeroen van den Hoven han propuesto que el sistema sea diseñado de tal forma que el resultado de sus operaciones pueda siempre remontarse hasta «al menos un humano en la cadena de diseño y operación[15]». Sin embargo, ¿qué pasa si hay mucha gente implicada (considérese de nuevo el problema de las muchas manos), y si no está claro quién es responsable?

¿LAS MATANZAS SELECTIVAS MEDIANTE DRONES SON GUERRA O ASESINATO?

En segundo lugar, antes de plantear la pregunta sobre la guerra justa, deberíamos preguntarnos si la lucha con drones se considera siquiera guerra. En particular, no está claro si la matanza selectiva realizada por drones se considera guerra, o si es un asesinato o una ejecución. Por ejemplo, en enero de 2020, un dron estadounidense lo seleccionó como objetivo y mató al general iraní Qasem Soleimani cerca del aeropuerto internacional de Bagdad. Muchos comentaristas lo consideraron un asesinato. Está claro que esto es una matanza selectiva. Pero ¿es un acto de guerra? Si a los individuos se les da un nombre (se les coloca en una «lista de individuos que matar»), son seleccionados como individuos (como algo opuesto a ser parte de un «enemigo» indefinido y como una amenaza colectiva), y se les da muerte por parte de gobiernos sin un proceso legal, entonces tienen sentido las etiquetas como «asesinato» o «ejecución». La matanza selectiva es una forma de asesinato llevada a cabo por los gobiernos contra quienes percibe como enemigos. En virtud de esto, se identifica a los individuos, se les selecciona y se les mata fuera de un procedimiento judicial o del campo de batalla. Otro término es «ejecuciones extrajudiciales», que tiene un significado similar. Generalmente esta práctica es considerada ilegal bajo la ley internacional y moralmente incorrecta (en términos de la teoría moral, es este un argumento deontológico: está mal, independientemente de cualesquiera otras consideraciones [consecuencialistas]), a pesar de que algunos países han intentado justificarla apelando a la «guerra» contra el terrorismo, es decir, a la autodefensa contra los terroristas. El tema de la justificación de las matanzas selectivas se merece su propia discusión[16]. Con respecto a los robots usados para este tipo de matanzas, parece justificado adoptar un enfoque cauteloso (si no una prohibición) por buenas razones éticas y políticas. Christof Heyns, durante su tiempo en las Naciones Unidas en el revelador papel de ponente especial sobre ejecuciones extrajudiciales, sumarias o arbitrarias, ha planteado ciertas preocupaciones sobre la protección de la vida humana, así como la preservación de «un mínimo orden mundial» dado «el uso problemático y las justificaciones controvertidas a favor de los drones y de las matanzas selectivas». Defiende que la responsabilidad recae en aquellos que quieren desplegar robots autónomos letales para proporcionar «pruebas de peso» de que se permiten usos específicos en circunstancias específicas[17].

MATAR A DISTANCIA

En tercer lugar, matar a distancia puede considerarse moralmente problemático, ya que, psicológicamente hablando, es más fácil matar a distancia. Dave Grossman, antiguo profesor de West Point especializado en la psicología del matar en la guerra, ha sostenido que matar a alguien a corta distancia, como, por ejemplo, con un cuchillo o con las propias manos, es difícil porque existe una proximidad física y empática, y por el (miedo al) trauma posterior[18]. Cuando la distancia aumenta, como en el caso de la artillería o de los bombardeos, matar se vuelve más sencillo. La preocupación radica en que, con la lucha en la que intervienen drones, matar a distancia se convierte en algo semejante a un videojuego, haciendo que matar sea demasiado fácil[19]. El personal militar mantiene que la lucha con drones no es como un videojuego y que se usan procedimientos y tomas de decisión similares a los de las aeronaves tripuladas convencionales. Además, matar con drones implica no solamente que el piloto o la persona presione el botón: existe un equipo completo de operadores de drones implicado y, por supuesto, un comandante (presente o remotamente presente) que toma las decisiones. Dicho esto, la distancia creada y la mediación de tecnología específica (una pantalla o un *joystick*) continúan existiendo, y es probable que, al menos, tengan *algún* impacto sobre lo fácil que resulte matar. La preocupación viene de que la distancia física tiene también impacto en la distancia moral, de que, fenomenológicamente, la «lucha a través de pantallas» hace que desaparezcan la cara y el cuerpo de las personas y, por ello, las deshumaniza y elimina las barreras para matarlas[20]. Este razonamiento es aplicable no solo a los drones, sino a todo tipo de armas a distancia, como misiles y bombas, en donde hay una gran distancia entre los operadores y los objetivos. Piénsese sobre el lanzamiento de las bombas atómicas en Japón a finales de la Segunda Guerra Mundial: los que las lanzaron no vieron el daño y el sufrimiento concretos en tierra. No vieron la muerte y el sufrimiento de la gente ahí abajo. En su lugar, desde su posición como dioses arriba en el aire, vieron una nube con forma de champiñón y sintieron la onda de choque de la explosión. Esta distancia parece hacer que la empatía se vuelva imposible. Parte de este efecto de «distancia moral» también puede ocurrir en el caso de la lucha con drones a distancia.

> La preocupación viene de que la distancia física tiene también impacto en la distancia moral, de que, fenomenológicamente, la «lucha a través de pantallas» hace que desaparezcan la cara y el cuerpo de las personas y, por ello, las deshumaniza y elimina las barreras para matarlas.

De forma interesante y trágica, los operadores tienen sin embargo una visión mucho mejor de lo que ocurre en tierra y a las personas en la guerra de drones contemporánea. Ven lo que ellas están haciendo y cómo viven sus vidas. Considérese de nuevo la descripción que se encuentra al comienzo de este capítulo, o estas observaciones de otro exsoldado de drones:

> Siempre que leo comentarios de políticos defendiendo los programas *Unmanned Aerial Vehicle Predator y Reaper* (vehículo depredador aéreo no tripulado y segador), también conocidos como drones, desearía poder hacerles algunas preguntas. Comenzaría con: «¿Cuántas mujeres y niños han visto usted incinerar por un misil *Hellfire*?». Y: «¿Cuántos hombres han visto arrastrarse por un campo, intentando alcanzar el recinto más cercano en busca de ayuda mientras se desangraban por sus piernas cercenadas?».
>
> O incluso de forma más directa: «¿Cuántos soldados ha visto usted morir al otro lado de una carretera de Afganistán porque nuestros UAV, siempre tan precisos, fueron incapaces de detectar un IED [dispositivo explosivo improvisado] que esperaba a nuestro convoy?». […] Conocía los nombres de algunos de los jóvenes soldados que vi desangrarse hasta morir al lado de la carretera. Vi morir en Afganistán docenas de hombres en edad militar, en campos vacíos, en los bancos de los ríos, y algunos justo enfrente del recinto donde estaban esperando sus familias a que volvieran a casa de la mezquita[21].

Tales descripciones sugieren que la tecnología remota no dificulta, sino que hace posible las experiencias empáticas. Las cámaras y otras tecnologías conectadas al dron crean un nuevo tipo de «intimidad» a pesar de la distancia[22]. El «objetivo» aparece ahora como una esposa, un pariente, una

persona yendo a casa y así sucesivamente. El «objetivo» también tiene un nombre. La «intimidad distante» y la rehumanización a través del revestimiento y reencarnación del «objetivo» crea una carga psicológica por parte de aquellos que la experimentan[23]. Los operadores de drones involucrados en la acción letal informan de que experimentan estrés; no solo tienen que tomar decisiones duras (por ejemplo, ¿es este un grupo de mujeres y niños, o son combatientes?), sino que también ven lo que ocurre en el campo de batalla[24]. Al estar expuestos a la violencia gráfica de destrucción de hogares y restos de humanos, incluso estos «guerreros de drones» que luchan desde una distancia segura[25] se han sentido traumatizados. Teniendo en cuenta el argumento utilitarista de Müller, deben contabilizarse también *estas* consecuencias.

Esta intimidad distante, o lo que puede denominarse fenómenos de «proximidad remota», y sus consecuencias psicológicas no es solo una cosa mala: quizás pueda mitigar el problema de que sea más sencillo matar a distancia, ya que, en principio, podría hacer que matar fuera más difícil en lugar de más fácil. Sin embargo, es plausible que aún exista una diferencia significativa por lo que respecta a la experiencia entre ver una muerte a través de la pantalla y verla enfrente de tus ojos estando presente en el campo de batalla. Después de todo, la experiencia descrita por estos exsoldados no les hizo detenerse de forma inmediata y dejar de colaborar en las matanzas con drones. Continuaron realizando su oscuro deber. En cualquier caso, continúa habiendo una diferencia entre la guerra aérea y con «las botas en el suelo». Los soldados sobre el terreno se encuentran en contacto mucho más cercano con los lúgubres resultados de sus acciones letales, y esto funciona como una barrera psicológica contra la matanza indiscriminada, mientras que la tecnología de la lucha con drones al menos abre esta posibilidad: existe una distancia y (por lo tanto) no hay riesgo por parte del atacante. Como nos recordaba George Monbiot en uno de sus artículos de opinión, el estratega militar Carl von Clausewitz pensaba que normalmente la guerra no escala hasta la brutalidad absoluta debido al «riesgo al que se enfrentan las propias fuerzas». La preocupación es que, sin riesgo, hay menos contención a la hora de matar. Monbiot concluye que «con estas naves no tripuladas, los gobiernos pueden desarrollar una guerra de cobardes, una guerra de dioses[26]». Esto también me conduce a la siguiente preocupación: ¿es este tipo de «guerra» justa y virtuosa?

JUSTICIA Y VIRTUD

En cuarto lugar, se podría objetar si la relación de vigilancia y vulnerabilidad asimétrica creada por la distancia es justa. Los operadores de drones ven el «objetivo» sobre el terreno, pero la persona que constituye el objetivo del ataque puede no saber que está siendo vista y que es vulnerable a lo que quiera que decida hacer el operador del dron. El operador no está en peligro, mientras que el «objetivo» lo asume por completo. Esta situación de «*deus ex machina*», que le otorga al operador del dron una visión y unos poderes divinos, recuerda a la vigilancia en un panóptico. El panóptico alude al diseño de un edificio creado por el filósofo Jeremy Bentham en el siglo XVIII, y a menudo se hace referencia a él en las discusiones contemporáneas sobre el poder y la vigilancia. Cuando se usa en el diseño de una prisión, el panóptico permite que los guardias vean a los prisioneros, pero los prisioneros no pueden ver a los guardias: estos ni siquiera saben si les están viendo. Los drones crean una situación similar. Pero, en este caso, los guardias del panóptico también disparan. Esta situación asimétrica en términos de vigilancia, control y riesgo (que no se limita a los drones, sino que también ocurre con otras formas de lucha remota) se puede considerar una lucha injusta e, incluso, cobarde de parte de quienes despliegan los drones, que va en contra del principio ético de la justicia y contra el honor y la virtud militares. Christian Enemark ha defendido que el uso de drones libre de peligro desafía las nociones tradicionales de lo que implica ser un soldado, y plantea problemas con respecto a la virtud y el heroísmo. ¿Es tal guerra (aún) más virtuosa que el asesinato organizado[27]? El hecho de que muchas de las personas que perpetran estas guerras de drones puedan no estar preocupadas por la justicia o la virtud, y que puede haber claros intereses militares y comerciales en juego (por ejemplo, financiar a la industria armamentística), no cambia el carácter no ético de este uso de la tecnología: posiblemente empeore las cosas.

EXPLICABILIDAD Y RESPONSABILIDAD

En quinto lugar, si los drones utilizan IA para tomar sus propias decisiones, esto plantea el problema de que algunos tipos de aprendizaje automático sean inexplicables desde un punto de vista humano y, así, el comportamiento del dron es impredecible. Esta cuestión es problemática cuando se plantean preguntas antes, durante y después de matar. ¿Cómo es posible la acción militar responsable bajo estas circunstancias? El desafío de

la responsabilidad con respecto a las tecnologías de la automatización no es solo que (recogiendo de nuevo la afirmación de los Sparrow) sea difícil mantener que cualquier persona o cosa sea responsable cuando lo que está en cuestión son sistemas altamente automatizados como los drones, ya que las máquinas carecen de cualquier *agencia moral* y se hallan fuera de nuestro control (considérese de nuevo la brecha de responsabilidad mencionada anteriormente). El problema es, también, que los que podemos llamar *pacientes* morales humanos pueden pedir, con razón, explicaciones sobre por qué o cómo se han tomado las decisiones a aquellos que despliegan sistemas autónomos. Aquí, la noción de responsabilidad se entiende, de nuevo, de una forma relacional y como rendición de cuentas: la cuestión de la responsabilidad no trata solo sobre los agentes y sus acciones y conocimiento como tales, sino también sobre ser responsable y rendir cuentas *a* otros. Aquí, entre esos otros se incluyen civiles no combatientes como víctimas o víctimas potenciales de ataques de drones (y sus representantes políticos), y, quizás, incluso los combatientes enemigos, todos los cuales pueden exigir, con razón, respuestas cuando se mata a personas. Tienen el derecho a saber cómo se tomó la decisión. Pero en el caso de drones asesinos que operan con IA y otros sistemas tecnológicos militares altamente «inteligentes» así como automatizados, ni la máquina ni el oficial al mando pueden ser capaces de ofrecer respuestas satisfactorias: la primera porque no puede responder en un sentido humano (responder como otro ser humano y ofrecer el tipo de respuestas que necesita un humano), el segundo porque no sabe cómo llegó la máquina a tomar su decisión. Si este es el caso, entonces parece más seguro limitar los grados de la automatización y mantener a los humanos en, o sobre, la cadena en sistemas que entienden: que *ellos* tomen o supervisen las decisiones y, por lo tanto, que asuman la responsabilidad entendida como la capacidad de *responder* por lo que hacen. Necesitamos control humano significativo, pero también control humano efectivo, y sistemas capaces de dar explicaciones y que permitan responder y rendir cuentas de una forma significativa.

¿Deberíamos permitir que las máquinas maten a seres humanos?

Finalmente, incluso si existe tal cosa como una guerra justa e incluso si no afloraran los problemas anteriores, se podría defender que es moralmente problemático en sí mismo tener máquinas que maten seres humanos. Se

podría plantear una exigencia deontológica directa al efecto de que los robots nunca deberían matar seres humanos, recurriendo a la primera ley de la robótica de Asimov que establece que *los robots no pueden hacer daño a los seres humanos*. Pero también se podrían ofrecer las siguientes tres razones que están relacionadas con las discusiones sobre lo que pueden y no pueden hacer las máquinas inteligentes. En primer lugar, las máquinas no solo pueden ser menos buenas que los humanos, *técnicamente* hablando, a la hora de discriminar entre objetivos y no objetivos, combatientes y no combatientes, etc. (que es, de nuevo, un problema *ius in bello*); como se ha dicho, las decisiones y las acciones de la máquina pueden ser *impredecibles*. ¿Queremos que haya decisiones impredecibles en lo que respecta a la vida y la muerte? Precisamente, esto fue, entre otras cosas, lo que querían evitar las perspectivas del tipo de Arkin mediante seres humanos «poco fiables» sustituidos por máquinas. Quizás no se pueda confiar en los humanos, pero las máquinas avanzadas que usan la LA pueden ser poco fiables en el sentido de impredecibles. Por supuesto, esta es una cuestión empírica: puede que cambie en el futuro. En segundo lugar, incluso si fueran lo suficientemente buenas tecnológicamente hablando o, incluso, superiores a los humanos, estas máquinas carecerían aún de un grado suficiente de *agencia* moral. En particular, carecerían de la capacidad de tomar decisiones éticas sobre la vida y la muerte. Por ejemplo, se podría mantener que la falta de emociones no es una ventaja, como afirman Arkin y Müller, sino una desventaja: si las emociones son parte de un buen juicio ético humano y pueden contribuir a una acción más ética, entonces necesitamos emociones y, por lo tanto, necesitamos mantener al humano en la cadena. La empatía, por ejemplo, puede proteger contra muertes sin control al estilo de videojuego, si esto es siquiera una adecuada manera de describirlo. En tercer lugar, incluso si las máquinas tuvieran algún tipo de moralidad de máquina e incluso si no hubiera ninguna objeción de principio a este tipo de moralidad, carecen, sin embargo, de paciencia moral de tipo humano y, por lo tanto, no se les debería permitir matar a seres humanos. Los humanos saben lo que significa tener miedo a la muerte y experimentar su vulnerabilidad. Si una entidad no está viva y carece de la experiencia de la existencia y vulnerabilidad humanas, entonces probablemente no debería permitírsele quitar la vida de ningún ser humano o tomar decisiones sobre la vida y la muerte de ninguno de ellos.

> Si una entidad no está viva y carece de la experiencia de la existencia y vulnerabilidad humanas, entonces probablemente no debería permitírsele quitar la vida de ningún ser humano o tomar decisiones sobre la vida y la muerte de ninguno de ellos.

Tanto la falta de agencia moral como la de paciencia moral completas y similares a las humanas crean asimetrías fundamentales entre los robots y las máquinas, y constituyen buenos argumentos en contra de matar de forma totalmente automatizada[28]. Como señala Heyns en su informe de las Naciones Unidas, «las máquinas carecen de moralidad y de mortalidad, y deberían, como resultado, no tener poder sobre la vida y la muerte humanas[29]». Matar a seres humanos, incluso si está justificado, debería quedar en manos de los humanos.

Durante los últimos años y, en parte, pero en gran medida, como resultado del activismo y de la presión de ciertos grupos de la sociedad civil y de expertos, existe un apoyo creciente hacia esta posición a nivel internacional. El secretario general de las Naciones Unidas Antonio Guterres dijo en una reunión de expertos en sistemas de armas autónomas letales que «las máquinas con el poder y el criterio para seleccionar objetivos y eliminar vidas sin implicación humana son políticamente inaceptables, moralmente repugnantes, y deberían estar prohibidas por leyes internacionales[30]». El Parlamento Europeo ha solicitado negociar una prohibición internacional. Por ejemplo, el Comité de Defensa del Parlamento belga ha adoptado una resolución que exige al Gobierno el apoyo a una prohibición internacional[31]. Algunas empresas y cada vez más trabajadores de la tecnología se niegan a desarrollar estos sistemas. Sin embargo, los principales actores militares y

políticos globales continúan con sus planes, y desarrollan y/o usan los UAV con capacidad letal.

Discutir la ética de los drones que matan nos hace reflexionar no solo sobre la tecnología autónoma, sino también sobre la justificación de la guerra, así como lo que hacen y deberían (o no) hacer en ella, matar a distancia, la virtud militar y la política global. Por el camino se han usado diferentes tipos de argumentos y teorías morales. Por ejemplo, Arkin y Müller usan un razonamiento consecuencialista para defender el uso de tales máquinas, mientras que muchos de los argumentos en contra se basan en perspectivas deontológicas sobre qué tipo de acciones son moralmente permisibles (e incluso había un argumento de la ética de la virtud sobre la cobardía). Al final, la discusión nos ha llevado a considerar una perspectiva filosófica del ser humano y de la existencia humana que toma en serio el significado moral de la mortalidad y la vulnerabilidad humanas para reflexionar sobre el hecho de matar. Es una perspectiva que, junto con todos los tipos de principios éticos y legales, tales como los principios bioéticos y los derechos humanos, podrían formar la base para una ética de los drones con capacidad de matar o, quizás se podría decir, una ética que no trate sobre las máquinas y *su* moralidad, sino que busque la protección de los *humanos* y de sus *vidas*.

Capítulo 8

Espejos robóticos más allá de lo humano: la ética de los robots como una ética medioambiental

Como organizadores de la conferencia Robophilosophy 2018 en Viena preguntamos a los participantes, ¿por qué construimos robots y por qué deberíamos construirlos[1]? Se ofrecieron muchas razones: construimos robots como medios para alcanzar fines económicos, realizar tareas sucias o peligrosas para nosotros, hacer avanzar nuestro conocimiento científico, satisfacer necesidades humanas como la socialización, proyectar y externalizar nuestras esperanzas y miedos, satisfacer nuestros deseos narcisistas, crear un mundo mejor o trascender la condición humana. Los robots también son parte de contextos y procesos históricos, sociales, culturales y lingüísticos. Siempre hay un «alguien» que construye robots (también nos tenemos que hacer preguntas sobre la sociedad, la política y el poder), y construimos robots con palabras, contamos historias sobre robots e, incluso, hablamos con robots. Además, los robots expresan normas y amenazan o ayudan a que nos demos cuenta de valores éticos[2]. Pero una de las razones conocidas más interesantes que se ofrecieron fue que los robots nos ayudan a comprender lo que nos hace humanos.

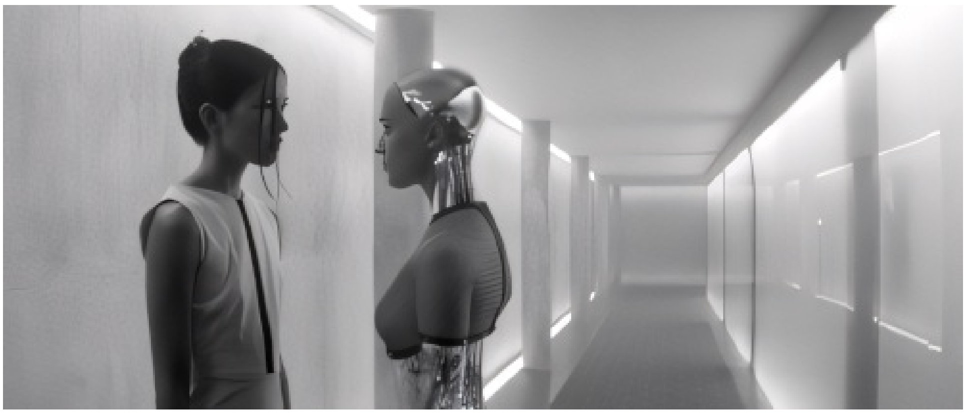


Figura 7. Fotograma de la película *Ex Machina*. Por cortesía de Universal Studios Licensing LLC.

Sobre esto es sobre lo que trata este capítulo. Las preguntas acerca de robots nos conducen a preguntas sobre humanos y la humanidad. ¿Qué implica usar a los robots como espejos de humanidad, qué podría ser posiblemente problemático en este proyecto y cuáles son las implicaciones para la ética de los robots?

Espejos robóticos: todo sobre los humanos

La ética de los robots y, de forma más general, la filosofía de la robótica no trata solo sobre robots, sino que también lo hace en gran medida sobre los humanos. Como muestran los capítulos de este libro, la ética de los robots trata, por supuesto, sobre máquinas, pero también se preocupa por un rango de temas que tratan todos ellos sobre lo humano: cuán significativo es el trabajo, cómo se están transformando nuestras sociedades y cómo deberían estar organizadas, qué son las buenas relaciones sociales y personales, si es aceptable engañar a las personas, qué implica respetar la dignidad humana, qué es un buen sistema sanitario, qué son la ética y la moral, qué queremos decir cuando hablamos de agencia y responsabilidad morales, si les debemos algo a los no humanos, cómo se fundamenta la posición moral de los propios humanos, qué es una guerra justa y aceptable moralmente (si es que es aceptable), y cómo la experiencia de la vulnerabilidad y la mortalidad humanas pueden operar como base para la ética. La ética de los robots trata, así, sobre nosotros: el presente y el futuro de nuestra moralidad, las sociedades y la existencia humana. En este sentido, los robots no son solo el objeto de nuestra ética y de otros trabajos filosóficos, como algo *sobre* lo que reflexionamos, pero que no nos concierne. Al contrario, los robots funcionan como espejos que *nos* muestran y reflejan, esto es, al ser humano en todas sus facetas, y con todos sus problemas y desafíos, incluyendo los éticos. La ética de los robots está siempre conectada con la antropología filosófica. Usamos los espejos robóticos como una herramienta filosófico-antropológica.

los robots funcionan como espejos que *nos* muestran y

reflejan, esto es, al ser humano
en todas sus facetas, y con todos
sus problemas y desafíos,
incluyendo los éticos.

La idea de usar las máquinas como espejos para la reflexión filosófica sobre lo humano no es nueva. René Descartes ya lo hizo cuando defendió que los animales eran máquinas complejas, mucho mejor hechas que nuestras máquinas porque estaban hechas por Dios, aun siendo «una mera máquina», si bien los humanos tenían características distintivas. Tenemos cuerpos animales, nuestros cuerpos son máquinas y, sin embargo, además estamos dotados de la razón. Una máquina puede proferir palabras, pero nosotros podemos dar respuestas llenas de sentido a cualquier cosa que se diga y usar la razón como un instrumento universal en todo tipo de situaciones.

Así, a pesar de que puede que no seamos capaces de encontrar la diferencia entre un animal y una máquina con forma de animal y con los órganos de un animal, siempre reconoceríamos a los humanos reales[3]. Dicho de forma más general, siempre hemos usado otras entidades para reflexionar sobre nosotros mismos y, de hecho, para definirnos como humanos y para distinguirnos de otras entidades. Piénsese en los animales, como en el argumento de Descartes, pero también en dios(es), ángeles, demonios, gólems, homúnculos, máquinas, monstruos, zombis, alienígenas y, ciertamente, robots e IA.

Como muestra el ejemplo de la antropología de Descartes, estos ejercicios a menudo han constituido lo que he denominado «antropologías negativas»: tendemos a definir al ser humano en términos de lo que no es[4]. No somos *x*: no somos animales, no somos dioses y así sucesivamente. De forma similar, las *tecnoantropologías* que usan robots se formulan normalmente de forma negativa: no somos robots, no somos máquinas y no somos entidades con IA. Pero este enfoque negativo también implica que en la modernidad *necesitamos* aparentemente a los robots para definir a los humanos. Desempeñan un papel importante en nuestras antropologías negativas. Sean lo que que sean, son herramientas *tecnoantropológlcas*. Solo alcanzamos la

distinción mediante la construcción de un contraste. El espejo robótico nos ofrece una imagen negativa que nos permite delinear lo humano. Somos no-robots, no-máquinas. O somos más que robots, más que máquinas (como en Descartes).

Una influencia importante en esta tecnoantropología particular es el humanismo, entendido de forma amplia: una perspectiva filosófica que destaca los seres humanos y su valor. Las antropologías humanistas como se desarrollaron en Occidente se preocupan de distinguir a los humanos de los no humanos y, si es posible, mantener y defender esta distinción. Desarrollamos el test de Turing y otros test junto con experimentos mentales para asegurarnos de que podemos hacerlo. Otras culturas parecen menos obsesionadas con establecer y mantener tales distinciones. Por ejemplo, influenciada por las cosmovisiones sintoístas y budistas tradicionales, la cultura japonesa parece preocuparse mucho menos por la cuestión de qué hace especiales a los humanos.

En la ética de los robots, esta tecnoantropología negativa y el humanismo toman la forma, normativamente hablando, de defender lo humano frente a la máquina. Se contempla a los humanos y a los robots en un marco antagonista. La relación entre los humanos y los robots se define por una narrativa competitiva. Decimos que no deberían asumir nuestros trabajos. Es el humano, o la máquina. Los robots son nuestras herramientas: no deberían convertirse en nuestros amos. Los valores humanos deberían prevalecer. Basándose en una antropología que intenta delinear claramente lo humano de lo no humano y que coloca primero (o más arriba) lo humano, la ética (y, por lo tanto, la ética de los robots) es humanocéntrica y defensiva cuando se trata de lo que se percibe como una amenaza. Se considera que los robots nos están desafiando a un duelo por nuestro carácter distintivo y nuestra posición. Quizás ya estén atacando. Pero este enfoque humanocéntrico y defensivo no es la única opción hablando en términos antropológicos y éticos.

Más allá de lo meramente humano

¿Qué le ocurre a la ética de los robots si vamos más allá de una antropología y una ética humanocéntricas y negativas? ¿Deberían los espejos convertirse en ventanas abiertas y, si así fuera, cómo y cuáles son las consecuencias para la ética de los robots? Ya he mencionado esta cuestión en el capítulo 3, pero se merece una discusión más profunda. Existen al menos las siguientes opciones para una ética de los robots y, de forma más general,

para una antropología y una ética que vayan más allá de lo meramente humano.

En primer lugar, existe la opción a no resistirse a «las máquinas» y, en su lugar, aceptar que los humanos son *un tipo de robot.* Somos los robots, como proclama la famosa canción de Kraftwerk. En esta opción aceptamos que somos máquinas, aunque sean biológicas. Esta antropología científica moderna de tipo cartesiano implica que también podemos mejorarnos, como ya lo hacemos, mediante la medicina y las prótesis: ya somos *cyborgs*. La naturaleza nos hizo máquinas limitadas: podemos mejorarnos y actualizarnos, esto es, si es que somos necesarios en este universo a largo plazo. Quizás sea mejor que seamos reemplazados por máquinas altamente inteligentes: los robots con IA que sean más listos que nosotros. Podemos crear máquinas que sean (o se conviertan) en mejores que nosotros. Esta es, en pocas palabras, al menos una versión influyente del transhumanismo. La resistencia contra las máquinas es fútil, ya que nosotros y nuestro humanismo ya están anticuados. A través de la modificación genética y de técnicas como la edición genética, pero también con la ayuda de todo tipo de tecnologías materiales, como interfaces cerebro-ordenador y partes corporales artificiales, podemos y deberíamos «mejorarnos». Desde esta perspectiva nos convertimos en un tipo de robot, *cyborgs* (combinaciones de materiales biológicos y no biológicos) y, quizás, en algún momento se nos acoja o se nos reemplace por lo que el robotista Hans Moravec ha denominado nuestros «hijos de la mente» (*mind children*). Si nuestras máquinas se vuelven tan complejas como los seres humanos, estas serán nuestros descendientes. Como afirma Moravec, «liberados del ritmo lento de la evolución biológica, los hijos de nuestra mente serán libres de enfrentarse a retos inmensos y desafíos fundamentales en el amplio universo[5]». En lugar de temer que los robots ocupen nuestro puesto, Moravec sugiere que deberíamos mirar hacia un futuro posbiológico maravilloso en el que podamos participar.

Pero ¿cómo llegaremos hasta allí, y cuál será nuestro lugar? Los transhumanistas contemporáneos tienen opiniones diversas al respecto, y algunos están más preocupados por los riesgos potenciales que otros. Ray Kurzweil ha defendido que nos convertiremos en *cyborgs* y, en última instancia, nosotros mismos nos subiremos a internet[6]. Podremos, entonces, vivir en una esfera digital y/o integrarnos en una máquina. Su concepción es generalmente optimista. Nick Bostrom escribe sobre máquinas que nos superan en inteligencia, pero también está preocupado por lo que esto pueda

implicar para nosotros los humanos: él plantea el problema de cómo controlar lo que haría una superinteligencia (el problema del control[7]). *Si* ocurriera este desarrollo, nuestro destino como especie dependería de las acciones de tal superinteligencia. Max Tegmark cree que habrá una explosión de inteligencia y que se extenderá la vida inteligente por el cosmos. Si esto ocurriera, nos veríamos humillados por máquinas siempre más inteligentes que nosotros. Reconoce riesgos, pero en la conclusión de su libro señala que podemos mejorar las perspectivas de un final feliz de la revolución actual «si podemos crear una sociedad humana más armoniosa caracterizada por la cooperación hacia metas comunes» y asumimos nuestra responsabilidad como «guardianes del futuro de la vida», conformando este futuro ahora y de forma conjunta[8].

A la luz de estas concepciones transhumanistas, hacer ética de los robots puede implicar construir máquinas que tengan moralidad, o que sean lo suficientemente inteligentes para comprender la moralidad por sí mismas, inventando quizás un nuevo tipo de moralidad que sea superior a la de los humanos, que se ven obstaculizados de forma innecesaria por limitaciones tales como las emociones. Se podría esperar entonces que las máquinas pudieran producir la mejor ética, la más racional. En lugar de centrarse siempre en la ética humana, como hace el humanismo, deberíamos crear las condiciones bajo las cuales pueda avanzar la moralidad de la máquina. Aquí la ética de los robots significa ética de la máquina. Con pensamiento racional y experimentos mentales como los juegos de moralidad, los filósofos pueden ayudar a establecer, junto con los científicos y los investigadores en robótica, una ética de este tipo. Además, los filósofos pueden servir de ayuda reflexionando sobre el problema del control y otras implicaciones para el futuro de la humanidad. La ética de los robots implica, entonces, guiar al desarrollo de las máquinas de forma que beneficie a la humanidad. Por ejemplo, Bostrom discute en *Superintelligence* (Superinteligencia) algunos métodos y técnicas que podríamos utilizar para evitar una catástrofe existential. Y la sugerencia de Tegmark se podría interpretar como un llamamiento a favor de asegurar que el desarrollo posterior de la robótica y de (otras) tecnologías inteligentes esté integrado en los esfuerzos que buscan transformar nuestra sociedad en una dirección armoniosa y cooperativa.

En segundo lugar, se podría tomar una ruta poshumanista que no se centre tanto en las máquinas nuevas excitantes y superinteligentes que podamos construir y en qué riesgos existenciales puedan suponer para la humanidad,

sino en aceptar a los robots como parte de su exaltación de la hibridación, el cruce de fronteras y la promiscuidad en relación con otras entidades. En esta concepción los humanos, los animales y los robots pueden y deben vivir juntos. En lugar de reemplazar a los humanos por robots, como teme el humanismo y espera el transhumanismo, la idea es que los humanos y los no humanos como los robots se puedan relacionar y fusionar. Los robots pueden ser nuestros otros artificiales amistosos, o podemos explorar la hibridación con las máquinas. Nos podemos acercar a ellos o, quizás, siempre hemos estado cerca. Qué «es» lo humano (no existe una «esencia» humana) no es ni debería estar claramente delineado. Debemos ir más allá de las categorías dualistas y binarias. Los *cyborgs* vuelven a aparecer, pero esta vez no tanto como combinaciones materiales de máquinas biológicas y no biológicas que deban mejorarnos como parte de un proyecto hipermoderno, sino como símbolos del cruce de fronteras posmoderno y de la liberación de lo binario. Las máquinas no son entonces inmigrantes inoportunos que quieren quitarnos nuestro lugar (como teme el humanismo), o que dan la bienvenida a los hijos que tenemos que liberar en el cosmos cuando se conviertan en mucho más inteligentes que nosotros (como desea el transhumanismo). Son un compañero potencial para el juego posmoderno, así como un objeto sobre el que podemos proyectar nuestros deseos psicológicos, aspiraciones éticas y esperanzas políticas. Es el momento para el amor robótico.

Donna Haraway es una influencia importante en esta versión del poshumanismo. En su famoso «Manifiesto Cyborg», Haraway ofrece la metáfora del *cyborg* como una herramienta para cuestionar las fronteras entre los humanos, los animales y las máquinas, conduciéndonos hasta un mundo en el que ya no tenemos miedo del «parentesco conjunto» con los animales y las máquinas, y en el que aceptamos a los monstruos que derriban barreras. No deberíamos temer los emparejamientos entre organismos y máquinas, y si los robots parecen estar vivos, este es solo un problema si nos aferramos un dualismo occidental y cartesiano[9]. Los robots, en la forma de robots que parecen estar vivos o como *cyborgs*, podrían desempeñar el papel de liberarnos de fronteras innecesarias que aún siguen vigentes en nuestro pensamiento y acción modernos. Pueden unirse y ayudar a establecer una utopía posmoderna en la que las concepciones binarias han desaparecido y donde exista la simbiosis: una convivencia de humanos, animales y máquinas. O quizás ya vivamos conjuntamente, *à la* Bruno Latour, en un mundo no moderno, en el que lo social consiste en humanos y no humanos, y donde hay

hibridación en todas partes[10]. En tal concepción no existe la necesidad de fantasías poshumanistas, ya que, para empezar, nunca hemos llegado a ser modernos. La filosofía moderna nos exige distinguir entre sujetos y objetos, humanos y máquinas, personas y cosas. Pero siempre hemos vivido en la hibridación. Los robots son ya parte de lo social.

Rosi Braidotti es otra voz contemporánea del poshumanismo. También de una forma posmoderna, interpreta lo poshumano en términos de identidades múltiples y defiende una reflexión posantropocéntrica. Como Haraway, cree que la distinción categórica entre los humanos y otras especies debería eliminarse. Braidotti discute también las fronteras difusas entre humano/máquina. En *Lo poshumano*, argumenta en contra de la distinción categórica entre naturaleza y cultura, y propone «una comprensión no dualista de la interacción naturaleza-cultura» dado que la ciencia y la tecnología han difuminado las fronteras. La complejidad de nuestras tecnologías cada vez más inteligentes y autónomas reside, para ella, en el núcleo del giro posantropocéntrico. Influenciada por el filósofo Félix Guattari (quien se basa en el biólogo Francisco Varela), Braidotti afirma que tanto los organismos biológicos como las máquinas son autopoiéticos (se autoorganizan). Como ya observó Haraway, las máquinas se vuelven más vivas, inmateriales, y los cuerpos se vuelven entidades conectadas a las máquinas. De nuevo nos encontramos con el *cyborg*, pero Braidotti incluye también «las masas anónimas de los malpagados, el proletariado digital que alimenta la economía global dirigida por la tecnología» en su poshumanismo imaginario, y defiende reflexionar en términos de una «relacionalidad radical» que funde el sujeto con su «entorno planetario tecnológicamente mediado», estableciendo un nuevo tipo de «relación simbiótica» y ética de interdependencia mutua. Sin embargo, rechaza «las formas extremas de poshumanismo dirigido por la ciencia que rechaza por completo la necesidad de un sujeto»: esto se puede interpretar como un intento de mantener la distancia frente a (algunas formas de) transhumanismo[11].

Desde un punto de vista poshumanisma, la ética de los robots podría, entonces, implicar que hemos encontrado una buena forma de relacionarnos y de vivir junto con las máquinas; y de desarrollar, usar y reflexionar sobre los robots de una forma que contribuya a derribar el pensamiento y las prácticas dualistas y binarias, y aceptar la hibridación y los monstruos. Una ética así no defiende lo humano frente a la máquina, sino que reflexiona sobre cómo dar forma a los acoplamientos humano-robot y construir una sociedad híbrida

humana/no humana. El arte y la ficción nos pueden ayudar a explorar este camino. Además, inspirados por el poshumanismo crítico de Braidotti, podríamos ir más allá de la ética de los robots entendida como la ética de la interacción humano-robot y considerar la economía política del *cyborg*. ¿Podrían los robots y la robótica ser parte de un sistema socioeconómico transformado que deje de ser explotador? ¿Qué implicaría considerar a los robots como parte de un entorno planetario tecnológicamente mediado? ¿Cómo puede la robótica ayudar a alcanzar una nueva simbiosis poshumanista entre humano y no humano, naturaleza y cultura, tecnología y biología?

En tercer lugar y finalmente, ir más allá de lo humano puede apuntar en la dirección del entorno natural y sugerir una «antropología» y una ética más ecológica y menos centrada en lo humano. A pesar de que este enfoque podría enlazar en principio con los anteriores (por ejemplo, con Latour o Braidotti), en esta opción el discurso es menos tecnófilo que en el transhumanismo o en el poshumanismo. El foco se coloca, antropológica y éticamente hablando, en el entorno natural, el planeta o la Tierra. Se podría entonces defender que, en lugar de vernos hipnotizados por la ciencia ficción transhumanista y las fantasías poshumanistas sobre *cyborgs*, deberíamos centrarnos en los problemas reales y urgentes del entorno natural y del planeta como el cambio climático. Si en realidad queremos hacer alguna mejora, no debemos olvidar que dependemos por completo de la Tierra y de la naturaleza para cualquier desarrollo tecnológico: mejor que intentemos mejorar *esa* relación y *esas* condiciones. Y si queremos ser más relaciónales, deberíamos comenzar por preocuparnos por el entorno natural, por las ecologías de las que somos una parte.

En lugar de vernos hipnotizados por la ciencia ficción transhumanista y las fantasías poshumanistas sobre *cyborgs*, deberíamos centrarnos en los problemas reales y urgentes del

entorno natural y del planeta
como el cambio climático.

En esta concepción, la ética de los robots no debería preocuparse solamente por los humanos (como hacen los humanistas) o los robots (como en las concepciones transhumanistas y poshumanistas). El desarrollo y el uso de los robots, si es que se necesita, debería contribuir a la consecución de un entorno natural y un planeta que sean ecológicamente sostenibles, y no estén socavados o destruidos por los humanos y sus tecnologías.

Esto es importante no solo porque otros seres vivos y entornos naturales tengan un valor inherente, además de los seres humanos: incluso desde un punto de vista antropocéntrico deberíamos trabajar en pos de un mundo sostenible, incluyendo un clima habitable en la Tierra. Bien puede que sea la única forma de *sobrevivir,* y es ciertamente la única forma de (comenzar o continuar a) vivir bien y de florecer. Además, el humanismo no es el único culpable. *Tampoco* los esfuerzos transhumanistas y científicos modernos por manipular la naturaleza y escapar del planeta *ni* las concepciones poshumanistas de la hibridación naturaleza-cultura son suficientes para asegurar que existe un futuro para los humanos y otros seres vivos. Si la robótica y la ética de los robots continúan su obsesión con las metas y valores humanos, o permanecen centradas solamente en los sueños de un futuro tecnológicamente perfecto (con o más allá de lo humano), contribuirán a la corrupción y destrucción del ecosistema(s) del único planeta que tenemos (la Tierra) y, por tanto, en última estancia, a la extinción de muchas especies incluyendo, quizás, la humana. Por el contrario, necesitamos una ética de los robots que esté centrada en el medioambiente. Para concluir, en la ética de los robots necesitamos no solo prestar atención a cómo pueden ayudarnos «los robots medioambientales» a responder a los «cambios medioambientales sin precedentes debidos a la influencia humana en los sistemas terrestres», así como a las cuestiones políticas y éticas que plantean[12]. La propia ética de los robots debería transformarse y convertirse en una *ética medioambiental.*

¿Qué quiere decir esto? ¿Qué implica concebir la ética de los robots como una ética medioambiental? Propongo distinguir entre una interpretación estrecha (o débil) y una amplia (o fuerte). La interpretación débil afirma que

necesitamos que los robots sean medioambientalmente amistosos en el sentido de que, dicho de forma negativa, minimicen o (preferentemente) eviten causar daño al entorno natural y a sus seres vivos y que, de forma positiva, nos ayuden a responder ante los desafíos medioambientales como el cambio climático (estos son los robots medioambientales que reclaman Aimee van Wynsberghe y Justin Donhauser). Alcanzar estos objetivos en y con la robótica ya sería un enorme paso adelante, considerado desde un punto de vista ambientalista y ecologista. Espero que este objetivo reciba, o esté ya recibiendo, mucho apoyo de los miembros de la comunidad robótica y más allá de ellos. Durante las últimas décadas se han vuelvo más o menos corrientes ciertas formas modestas de pensamiento medioambiental y ecológico. Muchos investigadores ya están interesados, o trabajando, en robótica medioambientalmente amistosa y sostenible. La interpretación fuerte, por el contrario, pone en duda el concepto mismo de «robot» y sus futuros tecnológicos. ¿Qué tipo de criaturas artificiales necesitamos, caso de que las necesitemos, dada la situación medioambiental en la que nos encontramos actualmente? ¿O los necesitamos porque lo humano debería sustituirse? ¿Cuál es el futuro de la robótica a la luz de una concepción radicalmente relacional y medioambiental?

Las respuestas a estas cuestiones pueden depender de qué posición se tome en relación a los otros enfoques mencionados. Un transhumanista, si está inspirado por el ambientalismo, y no defiende abandonar la Tierra y colonizar otros planetas, puede defender que se mejore a los humanos y que sean finalmente sustituidos por, o fusionados con, otras criaturas que muy probablemente hagan un mejor trabajo a la hora de gestionar el planeta. En este escenario los robots se pondrán al frente para alcanzar un objetivo ambientalista: las máquinas superinteligentes salvarán el planeta, a pesar de que pueda no merecer la pena salvar a la humanidad tal como es. Los poshumanistas críticos y los ambientalistas irán en otras direcciones, acogiendo quizás a las criaturas artificiales junto con las naturales e invitándolas a cohabitar esta tierra, o cuestionando las suposiciones modernas y transhumanistas, como el antropocentrismo y sus consecuencias para el medioambiente y el planeta.

He aquí *mis* preguntas, inspiradas por los últimos enfoques. ¿Podría ser que comprender la ética de los robots como una ética medioambiental tenga que conducir a abandonar el proyecto mismo de la robótica moderna? Si los robots son herramientas para obtener el control total sobre la Tierra, y si esto

es problemático, ¿es entonces suficiente construir robots medioambientalmente amistosos, o necesitamos cambios más radicales que quizás no impliquen de ningún modo a los «robots»? Si los robots son esclavos para que nos sirvan, pero si las relaciones sociales y los patrones de esclavitud, dominación y las relaciones hegemónicas en las que se encuentran integradas son moralmente reprehensibles, ¿necesitamos entonces a los «robots»? Si los robots reflejan lo humano, pero ese «humano» asume la cara espantosa de la subjetividad hegemónica junto con el control totalitarista de las personas y de la naturaleza, ¿necesitamos este tipo de espejo? La literatura poshumanista y ambientalista ofrece sugerencias a favor de caminos alternativos para la reflexión ética y política: formas de pensamiento que van más allá del humanismo anticuado y que rechazan las perspectivas modernas de la naturaleza de un antropocentrismo y dualismo fuertes. También ofrecen enfoques que van en las nuevas direcciones *políticas*, cuestionando las estructuras modernas de poder. Se necesita profundizar más sobre la política de la robótica.

A pesar de que está lejos de quedar claro lo que implican para el futuro de la tecnología y de la robótica estos enfoques y esta interpretación fuerte de la ética de los robots como ética medioambiental, me parece que una ética de los robots crítica y filosófica no debería evitar explorar estas direcciones, y preguntarse por estas cuestiones y discutirlas. Por ejemplo, si el diseño con valor sensible es el tipo de dirección correcta para la robótica, entonces ¿por qué tomar en cuenta solo los valores humanos? ¿Por qué debería centrarse el diseño en lo humano? ¿Es suficiente tener en cuenta *también* los valores de los no humanos y principios tales como la sostenibilidad, o podemos concebir un tipo de diseño más radicalmente medioambiental y políticamente crítico? Y, siguiendo con nuestro tema, ¿el resultado de tal diseño seguiría contando como un *robot*? ¿Cómo sería o a qué se parecería un «posrobot» medioambiental?

> Si los robots reflejan lo humano,
> pero ese «humano» asume la
> cara espantosa de la subjetividad

hegemónica junto con el control totalitarista de las personas y de la naturaleza, ¿necesitamos este tipo de espejo?

Los bocetos de las posiciones filosóficas de este capítulo son algo caricaturescos, pero pretenden sacar a la luz de forma más clara las diferencias entre las diversas direcciones (humanista, transhumanista, poshumanista y medioambiental) e invitar a los lectores a reflexionar sobre las posibilidades que pueden ofrecer estos enfoques para el futuro desarrollo de la ética de los robots. De nuevo, vemos cómo la ética de los robots no trata solo sobre robots, máquinas o artefactos: se preocupa por las «grandes» preguntas sobre los humanos y (hasta el punto en que necesitamos, o podemos, ir más allá de los humanos en nuestra reflexión y en la ética) en cualquier caso, las grandes cuestiones *para* los humanos que solo pueden ser respondidas *por* humanos. Los robots pueden ayudarnos en este ejercicio como herramientas y un espejo para nuestra reflexión sobre estas cuestiones. Como hemos visto, el espejo muestra muchas imágenes distintas. Pero, al final, los humanos tenemos que interpretar, imaginar, negociar y decidir como personas, sociedades y, quizás, como humanidad que vive en este planeta, qué dirección seguimos y debemos seguir. En otras palabras, nosotros los humanos tenemos que reflexionar sobre los robots, pero también sobre las posiblemente más importantes cuestiones planteadas por los robots, la robótica y la interacción humano-robot. Esta reflexión no puede ni debe dejarse en manos de las máquinas. Ética y políticamente hablando, el humanismo puede ser problemático, al menos en los sentidos criticados por el transhumanismo y el poshumanismo. Sin embargo, epistemológicamente tiene sentido si implica que reconozcamos que siempre tenemos que pasar a través de lo humano y de la subjetividad humana cuando reflexionamos sobre los robots, la ética de los robots y otras cuestiones. No «usamos» simplemente el espejo, somos parte de la actividad de *reflejar*.

Finalmente, debería preguntarse *qué* humanos o *quién* (individuos, grupos o categorías de personas) deberían implicarse en la ética de los robots y en

esos ejercicios de reflejos, aparte de los académicos que hacen ética de los robots y filosofía robótica de forma profesional. Una respuesta es los robotistas. Como he sugerido en el primer capítulo, los ingenieros, desarrolladores y cualquiera implicado en el diseño y la creación de robots tienen una responsabilidad significativa con respecto a sus creaciones. Para permitirles ejercer esta responsabilidad y participar en la reflexión ética, es vital integrar la ética en el plan de estudios de los programas de educación técnica y de formación, no como complemento, sino integrada en todas las materias técnicas relevantes. Se deberían también realizar esfuerzos por crear más diversidad (por ejemplo, en términos de género) con respecto a la población estudiantil y a la composición de equipos que realizan investigación robótica. Además, en línea con ideas sobre la investigación e innovación responsables y el análisis de la responsabilidad en términos de las muchas manos, deberían participar muchas otras personas a la hora de hacer ética de los robots, como los usuarios de los robots, por supuesto, pero también otras partes interesadas, como los inversores y los accionistas, los ejecutivos y otros gestores de empresas robóticas, los consultores, los profesionales y facultativos de un campo particular (por ejemplo, la sanidad o la educación), los gobiernos y las organizaciones no gubernamentales, como las organizaciones de derechos de los consumidores y los sindicatos. Si toman parte, o influyen, en las decisiones sobre la robótica, o cargan con las consecuencias de la robótica, también deberían reflexionar sobre su ética y política. Si es cierto que, al final, la ética de los robots (como cualquier ética de la tecnología) nos preocupa a todos como miembros de una comunidad, ciudadanos de una sociedad particular y miembros de la humanidad, si somos todos partes interesadas hasta cierto punto, entonces todos debemos reflexionar, opinar y aceptar alguna responsabilidad sobre el asunto. Dejemos que el espejo robótico nos muestre cuán democráticos somos realmente, cuánto podemos y queremos ser cuando se trata del futuro de la tecnología y de la sociedad.

Agradecimientos

Quisiera agradecer a mis editores en MIT Press, Philip Laughin, Alex Hoopes y Virginia Crossman, por apoyar el proyecto de este libro. Agradezco calurosamente a mi colega Zachary Storms por ayudarme a darle formato al manuscrito y organizar los *copyrights* para las imágenes. También agradezco a mis amigos y familia por estar ahí ayudándome: *online* y *offline*, y durante tiempos difíciles. Finalmente, estoy agradecido a mis estupendos amigos y colegas de la comunidad de la ética de los robots. Ha sido un placer participar en discusiones y pasar tiempo con ellos durante los últimos quince años. El campo de la ética de los robots no sería lo que es hoy sin ellos. ¡Continuemos haciendo un buen trabajo!

Glosario

Androides. Robots diseñados para parecerse a los humanos.

Antropomorfización. Atribuir características humanas a entidades no humanas, como los dioses, los animales u otros objetos. En la robótica y en la interacción humano-robot, el término se refiere al fenómeno en el que los humanos proyectan en un robot propiedades similares a las humanas, como los sentimientos.

Brecha de responsabilidad. En la ética de los robots y de la automatización, por «brecha de responsabilidad» se entiende el problema que plantea el hecho de que los robots y los sistemas autónomos han incrementado su autonomía y agencia, pero no su responsabilidad. La responsabilidad se puede adscribir entonces a los humanos, pero ¿quién es responsable? y ¿pueden aún los humanos intervenir a tiempo?

Capitalismo. Un sistema económico y político basado en la propiedad privada de los medios de producción y en la competición del libre mercado que, de acuerdo con Karl Marx y otros críticos, conduce a la acumulación de capital y del poder en manos de unos pocos, junto con la explotación de los trabajadores.

Cuarta revolución (industrial). La idea de que, tras las tres revoluciones industriales previas, cada una de las cuales fue posible mediante una nueva tecnología (por ejemplo, la máquina de vapor), la industria está siendo transformada ahora por el uso de robots autónomos e inteligentes, así como por (otros) sistemas ciberfísicos y el internet de las cosas.

Dentro de la cadena *versus* sobre la cadena *versus* fuera de la cadena. Distinciones que se refieren a cómo están implicados los humanos en el control de sistemas autónomos. ¿Tienen los humanos un control directo y se necesitan sus acciones para que el sistema actúe (dentro de la cadena), solo supervisan (sobre la cadena), o no hay interacción con el sistema, ya que funciona de forma autónoma (fuera de la cadena)?

Dilema del tranvía. Un experimento mental sobre ética que presenta el siguiente problema o alguna variante suya: un tranvía está circulando hacia cinco personas que están atadas a la vía, y tú controlas una palanca

que, si se tira, redirige al tranvía hacia otra vía donde hay una sola persona atada a ella. ¿No haces nada o tiras de la palanca?

Drones asesinos. UAV (ver más abajo) con capacidad letal, esto es, con la capacidad de matar.

Enfoque de las capacidades. Originalmente una teoría económica, se refiere ahora a un marco conceptual normativo que entiende la libertad de las personas para alcanzar el bienestar en términos de sus capacidades, esto es, lo que los individuos son capaces de hacer.

Ética. Un subcampo de la filosofía preocupado por cuestiones normativas sobre lo correcto (comportamiento) y lo bueno (vida).

Ética de los robots. Un campo de la filosofía de la tecnología y de la ética aplicada preocupado por las cuestiones normativas con respecto a los robots. Se puede referir a cómo deberían comportarse los humanos que usan y desarrollan la tecnología (por ejemplo, hacia otros humanos o robots), o cómo deberían comportarse los robots.

Experimento del Mago de Oz. Experimento de investigación en el que los sujetos interactúan con un ordenador o robot que creen que es autónomo e inteligente, pero que realmente está siendo operado (de forma remota) por un ser humano.

Fetiche. Término prestado del discurso freudiano en el que se hace referencia a una forma específica de patología sexual, se usa aquí en el sentido más general de un objeto (deseado) fabricado por los humanos y que es un sustituto de otra cosa distinta.

Filosofía robótica. La filosofía que se ocupa de las cuestiones sobre los robots.

Humanismo. Una perspectiva filosófica que coloca en el centro a los seres humanos y a su valor.

Industria 4.0. Usado originalmente en el contexto de la política industrial alemana, el término se refiere ahora a la idea de que estamos desplazándanos hacia una producción y hacia fábricas más inteligentes, usando el internet de las cosas y el análisis de *{big data)* datos junto con otras nuevas tecnologías, como la impresión 3D, la tecnología *cloud*, la realidad aumentada, etc.

Inteligencia artificial general. La inteligencia hipotética de una máquina que puede comprender o aprender cualquier tarea cognitiva que pueden realizar los humanos.

Investigación e innovación responsables. Procesos de investigación e innovación en los que las partes interesadas y los innovadores sociales son mutuamente responsables unos con otros, y en los que la ética está integrada desde el inicio.

Ius ad bellum versus ius in bello (**también**, ***ius ad bellum versus ius in bello***). La distinción entre la cuestión de la justificación para (comenzar) una guerra y la cuestión de la conducta justa en y durante la guerra.

La vida buena ***(eudaimonía)***. Un término filosófico. Los filósofos tienen diferentes opiniones sobre qué significa una vida buena. En la tradición aristotélica está conectada con alcanzar la excelencia personal o la virtud.

Leyes de la robótica de Asimov. Un conjunto de reglas éticas para los robots que ofrece el escritor de ciencia ficción Isaac Asimov. Las tres leyes son: un robot no puede hacer daño a un ser humano o, a través de la inacción, permitir que se le haga daño a un humano; un robot debe obedecer las órdenes dadas por seres humanos excepto si estas órdenes entran en conflicto con la primera ley; y un robot debe proteger su propia existencia en la medida en que tal protección no entre en conflicto con la primera o segunda ley. Más tarde, se añadió la ley cero: un robot no puede hacer daño a la humanidad o, por inacción, permitir que le hagan daño a esta.

Pacientes morales. Entidades hacia las que los agentes morales tienen responsabilidades morales. Se encuentran en la parte receptora de la relación moral y hay algo que se les debe.

Poshumanismo. Un abanico de perspectivas filosóficas que expande el círculo de la preocupación ética y filosófica por los no humanos, y evalúa de forma positiva el cruce de fronteras y las fusiones, como usar la metáfora del *cyborg*.

Renta básica universal. Un pago periódico realizado por el Estado a todos los ciudadanos de forma individual sin el requisito de trabajar o de informar de su actividad.

Robot. Un robot es una máquina autónoma capaz de sentir su entorno, realizar computaciones para tomar decisiones y realizar acciones en el mundo real (descripción basada en la definición del Institute of Electrical and Electronics Engineers). Este libro se centra en los robots que, además de esta definición, tienen *hardware*, así como un alto nivel de autonomía, inteligencia e interactividad.

Robots de asistencia social. Robots que pueden ayudar a las personas en contextos como la educación, el entrenamiento, la terapia y la

rehabilitación a través de interacciones sociales.

Robots medioambientales. Robots que nos ayudan a responder a los desafíos medioambientales, incluyendo el problema del cambio climático y sus consecuencias para los humanos y los ecosistemas de la Tierra.

Robots sociales. Robots diseñados para imitar el comportamiento social humano para facilitar la interacción humano-robot.

Sistemas ciberfísicos. Sistemas tecnológicos que integran procesos de computación, interconexión y físicos. El sistema es controlado por algoritmos y está conectado a internet.

Sociedad del ocio. La idea utópica de una sociedad en la que las máquinas han asumido los trabajos realizados previamente por los humanos, liberándolos de tal manera que disfruta de una vida de ocio.

Superinteligencia. Un agente hipotético que tiene una inteligencia que excede por mucho la del humano más inteligente.

Suspensión de la incredulidad. Suspender las propias facultades críticas y creer en algo por el mero hecho de la satisfacción epistémica.

Test de Turing. Un test propuesto por Alan Turing. El objetivo es comprobar si la máquina exhibe un comportamiento inteligente indistinguible del de los humanos. En el test, un evaluador humano juzga conversaciones en lenguaje natural entre un humano y una máquina diseñada para proporcionar respuestas similares a las humanas. Si el evaluador es incapaz de discernir de forma fiable entre la máquina y el humano, entonces la máquina ha superado el test.

Transhumanismo. La creencia de que la humanidad puede y debería evolucionar más allá de sus limitaciones actuales mediante la ciencia y la tecnología.

Valle de lo inquietante. Hipótesis propuesta por Masahiro Mori que afirma que la familiaridad se incrementa con el parecido humano, pero que, llegado un punto, existe un «valle de lo inquietante», en el que una sutil desviación de la apariencia humana crea un sentimiento de extrañeza o inquietud.

Vehículos aéreos no tripulados (UAV). Aeronave pilotada por control remoto o por ordenadores a bordo. Se podría llamar también «robot volador».

Bibliografía[*]

ABNEY, Keith, «Robotics, Ethical Theory, and Metaethics: A Guide for the Perplexed», en Patrick Lin, George Bekey y Keith Abney (eds.), *Robot Ethics: The Ethical and Social Implications of Robotics*, Cambridge, MA, MIT Press, 2012, págs. 35-52.

ANDERSON, Michael y ANDERSON, Susan (eds.), *Machine Ethics*, Cambridge, Cambridge University Press, 2011.

ARCHER, Joseph, «“Friendly” Hospital Robot Begins Trials to Help Stressed Nurses», *Telegraph*, 19 de septiembre, 2018. Disponible en: <https://www.telegraph.co.uk/tech nology/2018/09/19/friendly-hospital-robot-begins-trials-help-stressed-nurses/ >.

ARISTÓTELES, *Obras completas*, Madrid, Gredos, 2002.

ARKIN, Ronald, «The Case for Ethical Autonomy in Unmanned Systems», *Journal of Military Ethics*, 9, num. 4 (2010), págs. 332-341.

Article 36, «Killing by Machine: Key Issues for Understanding Meaningful Human Control», abril de 2015. Disponible en: <https://article36.org/ wp-content/uploads/2013/06 /KILLING_BY_MACHINE_6.4.15.pdf>.

ASARO, Peter, «What Should We Want from a Robot Ethic?», *International Review of Information Ethics*, 6 (2006), págs. 9-16.

—«How Just Could a Robot War Be?», en Philip Brey, Adam Briggle y Katinka Waelbers (eds.), *Current Issues in Computing and Philosophy*, Amsterdam, IOS Press, 2008, págs. 50-64.

ASIMOV, Isaac, «Runaround: A Short Story», *Astounding Science-Fiction*, 29, núm. 1, marzo de 1942, págs. 94-103.

AWAD, Edmond, DSOUZA, Sohan, KIM, Richard, SCHULZ, Jonathan, HENRICH, Joseph, SHARIFF, Azim, RAHWAN, Iyad y BONNEFON, Jean-François, «The Moral Machine Experiment», *Nature*, 563, 2018, págs. 59-64.

BARTNECK, Christoph, LÜTGE, Christoph, WAGNER, Alan R. y WELSH, Sean, *An Introduction to Ethics in Robotics and Al*, Cham, Suiza, Springer, 2021.

BEAUCHAMP, Tom y CHILDRESS, James, *Principles of Biomedical Ethics*, Oxford, Oxford University Press, 1994.

BENJAMIN, Ruha, *«Race after Technology»: Abolitionist Tools for the New Jim Code*, Cambridge, Reino Unido, Polity Press, 2019.

BODDINGTON, Paula, *Towards a Code of Ethics for Artificial Intelligence*, Cham, Suiza, Springer, 2017.

BORENSTEIN, Jason y Pearson, Yvette, «Robot Caregivers: Harbingers of Expanded Freedom for All?», *Ethics and Information Technology*, 12 (2010), 277-288.

BOSTROM, Nick, *Superintelligence: Paths, Dangers, Strategies*, Oxford, Oxford University Press, 2014.

BOTTING, Eileen Hunt, *Mary Shelley and the Rights of the Child*, Filadefia, University of Pennsylvania Press, 2018.

BRAIDOTTI, Rosi, *The Posthuman*, Cambridge, Reino Unido, Polity Press, 2013.

BRYNJOLFSSON, Erik y MCAFEE, Andrew, *The Second Machine Age: Work, Progress, and Prosperity in a Time of Brilliant Technologies*, Nueva York, W. W. Norton and Company, 2014.

BRYSON, Joanna, «Robots Should Be Slaves», en Yorick Wilks (ed.), *Close Engagements with Artificial Companions: Key Social, Psychological, Ethical and Design Issues*, Amsterdam, John Benjamins, 2010, págs. 63-74.

BUMILLER, Elisabeth, «A Day Job Waiting for a Kill Shot a World Away», *New York Times*, 30 de julio de 2012. Disponible en: <https://www.nytimes. com/2012/07/30/us/drone-pilots-waiting-for-a-kill-shot-7000-miles-away.html>.

BYRNE, Edmund, «Making Drones to Kill Civilians: Is It Ethical?», *Journal of Business Ethics*, 147 (2015), 81-93.

CALO, M. Ryan, «Robots and Privacy», en Patrick Lin, George Bekey y Keith Abney (eds.), *«Robot Ethics»: The Ethical and Social Implications of Robotics*, Cambridge, MA, MIT Press, 2012, págs. 187-201.

CAMPAÑA PARA DETENER A LOS ROBOTS ASESINOS, «Parliamentary Actions in Europe», 10 de julio de 2018. Disponible en: <https://www.stopkillerrobots.org/2018/07/parliaments-2/>.

CARPENTER, Julie, «The Existential Robot: Living with Robots May Teach Us to Be Better Humans», *Issues*, 108 (2014), 39-42.

—«Deus Sex Machina: Loving Robot Sex Workers and the Allure of an Insincere Kiss», en John Danaher y Neil McArthur (eds.), *Robot Sex:*

Social and Ethical Implications, 261-287, Cambridge, MA, MIT Press, 2017, págs. 261-287.

CARPENTER, Julie, DAVIS, Joan M., STEWART, Norah Erwin, Lee, Tiffany R., Bransford, John D. y Vye, Nancy, «Gender Representation and Humanoid Robots Designed for Domestic Use», *International Journal of Social Robotics*, 1 (2009), 261-265.

CLARK, Roger, «Asimov's Laws of Robotics: Implications for Information Technology», en Michael Anderson y Susan Anderson (eds.), *Machine Ethic*s, Cambridge, Cambridge University Press, 2011, págs. 254-284.

COECKELBERGH, Mark, «Virtual Moral Agency, Virtual Moral Responsibility», *Al and Society*, 24, núm. 2 (2009), 181-189.

—«Moral Appearances: Emotions, Robots, and Human Morality», *Ethics and Information Technology*, 12, núm. 3 (2010), 235-241.

—«Health Care, Capabilities, and Al Assistive Technologies», *Ethical Theory and Moral Practice*, 13 (2010), 181-190.

—«You, Robot: On the Linguistic Construction of Artificial Others», *Al and Society*, 26, núm. 1 (2011), 61-69.

—*Growing Moral Relations: Critique of Moral Status Ascription*, Nueva York, Palgrave Macmillan, 2012.

—«Care Robots, Virtual Virtue, and the Best Possible Life», en Philip Brey, Adam Briggle y Ed Spence (eds.), *The Good Life in a Technological Age*, 281-292, Nueva York, Routledge, 2012.

—«How I Learned to Love the Robot: Capabilities, Information Technologies, and Elderly Care», en Use Oosterlaken y Jeroen van den Hoven (eds.), *The Capability Approach, Technology and Design*, Dordrecht, Springer, 2012, págs. 77-86.

—«Drones, Information Technology, and Distance: Mapping the Moral Epistemology of Remote Fighting», *Ethics and Information Technology*, 15, núm. 2 (2013), 87-98.

—«E-Care as Craftsmanship: Virtuous Work, Skilled Engagement, and Information Technology in Health Care», *Medicine, Healthcare and Philosophy*, 16, núm. 4 (2013), 807-816.

—*Human Being @ Risk: Enhancement, Technology, and the Evaluation of Vulnerability Transformations*, Nueva York, Springer, 2013.

—«The Moral Standing of Machines: Towards a Relational and Non-Cartesian Moral Hermeneutics», *Philosophy and Technology*, 27, núm. 1 (2014), 61-77.

—«Robotic Appearances and Forms of Life: A Phenomenological-Hermeneutical Approach to the Relation between Robotics and Culture», en Michael Funk y Bernhard Irrgang (eds.), *Robotics in Germany and Japan: Philosophical and Technical Perspectives*, Berlin, Peter Lang, 2014, págs. 59-68.

—«Artificial Agents, Good Care, and Modernity», *Theoretical Medicine and Bioethics*, 36 (2015), 265-277.

—«Care Robots and the Future of ICT-Mediated Elderly Care: A Response to Doom Scenarios», *Al and Society*, 31, núm. 4 (2016), 455-462.

—«Drones, Morality, and Vulnerability: Two Arguments against Automated Killing», en Bart Custers (ed.), *The Future of Drone Use: Technologies, Opportunities and Privacy Issues*, The Hauge, T. M. C. Asser Press, 2016, págs. 229-237.

—*New Romantic Cyborgs*, Cambridge, MA, MIT Press, 2017.

—«How to Describe and Evaluate "Deception" Phenomena: Recasting the Metaphysics, Ethics, and Politics of ICTs in Terms of Magic and Performance and Taking a Relational and Narrative Turn», *Ethics and Information Technology*, 20, núm. 2 (2018), 71-85.

—«Why Care about Robots? Empathy, Moral Standing, and the Language of Suffering», *Kairos: Journal of Philosophy and Science*, 20, núm. 1 (2018), 141-158.

—*Moved by Machines: Performance Metaphors and Philosophy of Technology*, Nueva York, Roudedge, 2019.

—«AI, Responsibility Attribution, and a Relational Justification of Explainability», *Science and Engineering Ethics*, 26, núm. 4 (2020), 2051-2068.

—«Technology Games/Gender Games: From Wittgensteins Toolbox and Language Games to Gendered Robots and Biased Artificial Intelligence», en Janina Loh y Mark Coeckelbergh (eds.), *Feminist Philosophy of Technology*, Stuttgart, J. B. Metzler, 2020, págs. 27-38.

—«Should We Treat Teddy Bear 2.0 as a Kantian Dog? Four Arguments for the Indirect Moral Standing of Personal Social Robots, with Implications for Thinking about Animals and Humans», *Minds and Machines* (30 de diciembre de 2020). Disponible en: <https://doi.org/10.1007/sl 1023-020-09554-3>.

—*Ética de la Inteligencia Artificial*, Madrid, Cátedra, 2021.

—«How to Use Virtue Ethics for Thinking about the Moral Standing of Social Robots: A Relational Interpretation in Terms of Practices, Habits, and Performance», *International journal of Social Robotics,* 13, núm. 1 (2021), 31-40.

Coeckelbergh, Mark, Loh, Janina, Funk, Michael, Seibt, Johanna y Nørskov, Marco (eds.), *Envisioning Robots in Society: Power, Politics, and Public Space*, Actas de Robophilosophy 2018/TRANSOR 2018, University of Vienna, febrero de 2018, Amsterdam, IOC Press, 2018.

Coeckelbergh, Mark, Pop, Christina, Simut, Ramona, Peca, Andreea, Pintea, Sebastian, David, Daniel y Vanderborght, Bram, «A Survey of Expectations about the Role of Robots in Robot-Assisted Therapy for Children with ASD: Ethical Acceptability, Trust, Sociability, Appearance, and Attachment», *Science and Engineering Ethics*, 22, núm. 1 (2016), 47-65.

Danaher, John, *Automation and Utopia: Human Flourishing in a World without Work*, Cambridge, MA, Harvard University Press, 2019a.

—«Welcoming Robots into the Moral Circle: A Defense of Ethical Behaviourism», *Science and Engineering Ethics*, 26, núm. 4 (2019b), 2023-2049. Disponible en: <https://doi.org/10.1007/sll948-019-00119-x>.

Danaher, John y Mcarthur, Neil, *Robot Sex: Social and Ethical Implications*, Cambridge, MA, MIT Press, 2017.

Darling, Kate, «“Who’s Johnny?” Anthropomorphic Framing in Human-Robot Interaction, Integration, and Policy», en Patrick Lin, Keith Abney y Ryan Jenkins (eds.), *Robot Ethics 2.0: From Autonomous Cars to Artificial Intelligence*, Nueva York, Oxford University Press, 2017, págs. 173-188.

—«Extending Legal Protection to Social Robots», *IEEE Spectrum*, 10 de septiembre de 2012. Disponible en: <https://spectrum.ieee.org/extending-legal-protection-to-social-robots>.

Descartes, Rene, *Discurso del método*, Madrid, Alianza Editorial, 2011.

Dreyfus, Hubert L. y Dreyfus, Stuart E., «Towards a Phenomenology of Ethical Expertise», *Human Studies*, 14, núm. 4 (1991), 229-250.

Dreyfus, Stuart E. y Dreyfus, Hubert L., *A Five-Stage Model of the Mental Activities Involved in Direct Skill Acquisition*, Berkeley, Operations Research Center, University of California, 1980.

Duff, R. Antony, «Who Is Responsible, for What, to Whom?», *Ohio State Journal of Criminal Law*, 2 (2005), 441-461.

Enemark, Christian, *Armed Drones and the Ethics of War: Military Virtue in a Post-Heroic Age*, Abingdon, Reino Unido, Routledge, 2016.

Fischer, John M. y Ravizza, Mark, *Responsibility and Control: A Theory of Moral Responsibility*, Cambridge, Cambridge University Press, 1998.

Fletcher, Sarah R. y Webb, Philip, «Industrial Robot Ethics: The Challenges of Closer Human Collaboration in Future Manufacturing Systems», en Maria Isabel Aldinhas Ferreira, João Silva Sequeira, Mohammad O. Tokhi, Endre E. Kadar y Gurvinder S. Virk (eds.), *A World with Robots: International Conference on Robot Ethics: ICRE 2015*, Cham, Suiza, Springer, 2017, págs. 159-169.

Floridi, Luciano y Sanders, J. W., «On the Morality of Artificial Agents», *Minds and Machines*, 14, núm. 3 (2004), 349-379.

Flusser, Vilém, *Shape of Things: A Philosophy of Design*, Londres, Reaction Books, 1999.

Ford, Martin, *Rise of the Robots: Technology and the Threat of Jobless Future*, Nueva York, Basic Books, 2015.

Foro económico mundial, *The Future of Jobs Report 2018*, Geneva, World Economic Forum, 2018a. Disponible en: <http://www3.weforum.org/docs/WEF_Future_of_Jobs _2018.pdf>.

—*Towards a Reskilling Revolution: A Future of Jobs for All*, Génova, World Economic Forum, 2018b. Disponible en: <http://www3.weforum.org/ docs /WEF_FOW_Reskilling_Revolution.pdf>.

Fosch-villaronga, Eduard, *Robots, Healthcare, and the Law: Regulation Automation in Personal Care*, Abingdon, Reino Unido, Roudedge, 2020.

Foster, Malcolm, «Aging Japan: Robots May Have Role in Future of Elder Care», *Reuters*, 28 de marzo de 2018. Disponible en: <https://www.reuters.com/article/us-japan-ageing-robots-widerimage/aging-japan-robots-may-have-role-in-ftiture-of-elder-care-idUSKBN 1H33AB/>.

Frey, Carl Benedikt y Osborne, Michael A., «The Future of Employment: How Susceptible Are Jobs to Computerisation?», *Oxford Martin School Working Paper*, núm. 7, 2013. Disponible en: <https://www.oxfordmartin.Ox.ac.uk/ publications/the-future-of-employment/>.

Friedman, Batya, «Value-Sensitive Design», *Interactions*, 3, núm. 6 (1996), 16-23.

FRIEDMAN, Batya y HENDRY, David G., *Value Sensitive Design: Shaping Technology with Moral Imagination*, Cambridge, MA, MIT Press, 2019.

FUNK, Michael, SEIBT, Johanna y COECKELBERGH, Mark, «Why Do/Should We Build Robots? Summary of a Plenary Discussion Session», en Mark Coeckelbergh, Janina Loh, Michael Funk, Johanna Seibt y Marco Nørskov (eds.), *Envisioning Robots in Society: Power, Politics, and Public Space*, Proceedings of Robophilosophy 2018/TRANSOR 2018, University of Vienna, febrero de 2018, Amsterdam, IOC Press, 2018, páginas 369-384.

GROSS, Michael L., «Assassination and Targeted Killing: Law Enforcement, Execution or Self-Defence?», *Journal of Applied Philosophy*, 23, núm. 3 (2006), 323-334.

GROSSMAN, Dave, *On Killing: The Psychological Cost of Learning to Kill in War and Society*, Nueva York, Little, Brown and Company, 2009.

GUNKEL, David, *An Introduction to Communication and Artificial Intelligence*, Cambridge, Reino Unido, Polity Press, 2020.

—«The Other Question: Can and Should Robots Have Rights?», *Ethics and Information Technology*, 20, núm. 2 (2018), 87-99.

—*The Machine Question: Critical Perspectives on Al, Robots, and Ethics*, Cambridge, MA, MIT Press, 2012.

HARAWAY, Donna, «A Cyborg Manifesto», en David Bell y Barbara M. Kennedy (eds.), *The Cybercultures Reader*, Londres, Routledge, 2000, págs. 291-324.

HEIDEGGER, Martin, *The Question concerning Technology and Other Essays*, trad, de William Lovitt, Nueva York, Harper and Row, 1977.

HELVEKE, Alexander y NIDA-RÜMELIN, Julian, «Responsibility for Crashes of Autonomous Vehicles: An Ethical Analysis», *Science and Engineering Ethics*, 21, núm. 3 (2015), 619-630.

HEYNS, Christof, *Report of the Special Rapporteur on Extrajudicial, Summary or Arbitrary Executions*, Naciones Unidas, Consejo de Derechos Humanos, 9 de abril, 2013 (A/ HRC/23/47). Disponible en: <http://www.ohchr.org/Documents/HRBodies/HRCouncil/RegularSession/Session23/A-HRC-23-47 _en.pdf>.

HILDEBRANDT, Mireille, *Smart Technologies and the End(s) of Law: Novel Entanglements of Law and Technology*, Cheltenham, Reino Unido, Elgar, 2015.

HORSTMANN, Aike C., BOCK, Nikolai, LINHUBER, Eva, SZCZUKA, Jessica M., STRASSMANN, Carolin y KRÄMER, Nicole C., «Do a Robot's Social Skills

and Its Objection Discourage Interactants from Switching the Robot Off?», *PLoS ONE*, 13, num. 7 (2018), e0201581. Disponible en: <https:// doi.org/10.1371/journal.pone.0201581 >.

HOVEN, Jeroen van den, Blaauw, Martijn, Pieters, Wolte y Warmer, Martijn, «Privacy and Information Technology», en Edward N. Zalta (ed.), *Stanford Encyclopedia of Philosophy*, invierno de 2019. Disponible en: <https://plato.stanford.edu/entries /it-privacy/>.

ISAAC, Alistair M. C. y Bridewell, Will, «White Lies on Silver Tongues: Why Robots Need to Deceive (and How)», en Patrick Lin, Keith Abney y Ryan Jenkins (eds.), *Robot Ethics 2.0: From Autonomous Cars to Artificial Intelligence*, Nueva York, Oxford University Press, 2017, págs. 157-172.

ISHIGURO, Hiroshi, «Android Science: Toward a New Cross-Interdisciplinary Framework», *Scientific American*, 294, núm. 5 (2006), 32-34.

JOHNSON, Deborah, «Computer Systems: Moral Entities but Not Moral Agents», *Ethics and Information Technology*, 8, núm. 4 (2006), 195-204.

KANT, Immanuel, *Lecciones de ética*, Barcelona, Planeta, 2013.

KITWOOD, Tom, *Dementia Reconsidered: The Person Comes First*, Buckingham, Reino Unido, Open University Press, 1997.

KREPS, Sarah, «Ground the Drones? The Real Problem with Unmanned Aircraft», *Foreign Affairs*, 4 de diciembre de 2013. Disponible en: <https:// www.foreignaffairs.com/articles *120*13-12-04/ground-drones>.

KURZWEIL, Ray, *The Singularity Is Near: When Humans Transcend Biology*, Nueva York, Penguin, 2005.

LATOUR, Bruno, *Nunca fuimos modernos*, Buenos Aires, Siglo XXI, 2007.

LEOPOLD, Todd, «HitchBOT, the Hitchhiking Robot, Gets Beheaded in Philadelphia», CNN, 4 de agosto de 2015. Disponible en: <https://edition.cnn.com/2015/08/03/us/hitchbot-robot-beheaded-philadelphia-feat/ index.html>.

LEVIN, Sam y Wong, Julia Carrie, «Self-Driving Uber Kills Arizona Woman in First Fatal Crash Involving Pedestrian», *Guardian*, 19 de marzo de 2018. Disponible en: <https:// www.theguardian.com/technology/2018/mar/19/ uber-self-driving-car-kills-woman-arizona-tempe>.

LIAO, S. Matthew (ed.), *Ethics of Artificial Intelligence*, Nueva York, Oxford University Press, 2020.

LIN, Patrick, ABNEY, Keith y BEKEY, George, «Robot Ethics: Mapping the Issues for a Mechanized World», *Artificial Intelligence*, 175 (2011), 942-949.

LIN, Patrick, ABNEY, Keith y JENKINS, Ryan (eds.), *Robot Ethics 2.0: From Autonomous Cars to Artificial Intelligence*, Nueva York, Oxford University Press, 2017.

LIN, Patrick, BEKEY, George y ABNEY, Keith (eds.), *Robot Ethics: The Ethical and Social Implications of Robotics*, Cambridge, MA, MIT Press, 2012.

—*Autonomous Military Robotics: Risk, Ethics, and Design*, Informe encargado por la Oficina de Investigación Naval del Departamento de la Marina de los Estados Unidos, premiado #N00014-07-l-1152, San Luis Obispo, California Polytechnic State University, 2008.

LINEBAUGH, Heather, «I Worked on the U.S. Drone Program: The Public Should Know What Really Goes On», *Guardian*, 20 de diciembre de 2013. Disponible en: <https://www.theguardian.com/commentisfree/2013/dec/29/ drones-us-military>.

LOH, Wulf y LOH, Janina, «Autonomy and Responsibility in Hybrid Systems: The Example of Autonomous Cars», en Patrick Lin, Keith Abney y Ryan Jenkins (eds.), *Robot Ethics 2.0: From Autonomous Cars to Artificial Intelligence*, Nueva York, Oxford University Press, 2017, págs. 35-50.

MACDORMAN, Karl E e ISHIGURO, HIROSHI, «The Uncanny Advantage of Using Androids in Cognitive and Social Science Research», *Interaction Studies*, 7, núm. 3 (2006), 297-337.

MARX, Karl, *El capital*, Madrid, Akai, 2022.

MATSUZAKI, Hironori y LINDEMANN, Gesa, «The Autonomy-Safety-Paradox of Service Robotics in Europe and Japan: A Comparative Analysis», *Al and Society*, 31, num. 4 (2016), 501-517.

MATTHIAS, Andreas, «The Responsibility Gap: Ascribing Responsibility for the Actions of Learning Automata», *Ethics and Information Technology*, 6, núm. 3 (2004), 175-183.

MAVROFOROU, Anna, MICHALODIMITRAKIS, Emmanuel, HATZITHEOFILOU, Konstantinos y GIANNOUKAS, Athanasios, «Legal and Ethical Issues in Robotic Surgery», *International Union of Angiology*, 29, núm. 1 (2010), 75-79.

MCCAMMON, Sarah, «The Warfare May Be Remote but the Trauma Is Real», NPR, 24 de abril de 2017. Disponible en: <https://www.npr.org/2017/04/24/525413427/for-drone-pilots-warfare-may-be-remote-but-the-trauma-is-real?t= 1571811594970&t= 16147 00024482>.

MCKAY, Tom, «19 Climate Change Activists Arrested for Drone Protest against Heathrow Airport Expansion», *Gizmodo*, 14 de septiembre de

2019. Disponible en: <https://gizmodo.com/19-extinction-rebellion-activists-arrested-in-drone-pro-1838122386>.

McKinsey Global Institute, *Jobs Lost, Jobs Gained: Workforce Transitions in a Time of Automation*, Nueva York, McKinsey Global Institute, diciembre de 2017.

Monbiot, George, «With Its Deadly Drones, the U. S. Is Fighting a Coward's War», *Guardian*, 30 de enero de 2012. Disponible en: <https://www.the-guardian.com/commentisfree/2012/jan/30/deadly-drones-us-cowards-war>.

Moravec, Hans, *Mind Children: The Future of Robot and Human Intelligence*, Cambridge, MA, Harvard University Press, 1988.

Mori, Masahiro, «The Uncanny Valley», traducido por Karl F. MacDorman y Norri Kageki, *IEEE Robotics and Automation Magazine*, 19, núm. 2 (2012), 98-100.

Møller, Michael, «Secretary-General's Message to Meeting of the Group of Governmental Experts on Emerging Technologies in the Area of Lethal Autonomous Weapons Systems», Charla presentada en la oficina de las Naciones Unidas, Ginebra, 25 de marzo de 2019. Disponible en: <https://www.un.org/sg/en/content/sg/statement/2019-03-25/ secretary-generals-message-meeting-of-the-group-of-governmental-experts-emerging-techno logies-the-area-of-lethal-autonomous-weapons-systems>.

Müller, Vincent, «Autonomous Killer Robots Are Probably Good News», en Ezio Di Nucci y Filippo Santoni de Sio (eds.), *Drones and Responsibility: Legal, Philosophical and Sociotechnical Perspectives on Remotely Controlled Weapons*, Londres, Roudedge, 2016, págs. 67-81.

Nordrum, Amy, «Cyderdyne's HAL Exoskeleton Helps Patients Walk Again in First Treatments at U. S. Facility», *IEEE Spectrum*, 3 de enero de 2019. Disponible en: chttps://spectrum.ieee.org/cyberdynes-hal-medical-exoske leton-helps-patients-walk-again-at-first-us-facility>.

Nozick, Robert, *Anarquía, Estado y Utopía*, Sevilla, Innisfree, 2015.

Nussbaum, Martha, *Las mujeres y el desarrollo humano*, Barcelona, Herder, 2002.

—*Las fronteras de la justicia*, Barcelona, Paidós, 2012.

Nussbaum, Martha y Sen, Amartya, *La calidad de la vida*, México, Fondo de Cultura Económico, 1999.

Nyholm, Sven, *Humans and Robots: Ethics, Agency, and Anthropocentrism*, Londres, Rowman and Littlefield, 2020.

NYHOLM, Sven y SMIDS, Jilles, «The Ethics of Accident-Algorithms for SelfDriving Cars: An Applied Trolley Problem?», *Ethical Theory and Moral Practice*, 19, mim. 5 (2016), 1275-1289.

PARKE, Phoebe, «Is It Cruel to Kick a Robot Dog?», CNN, 13 de febrero de 2015. Disponible en: <https://edition.cnn.com/2015/02/13/tech/spot-robot-dog-google/index.html>.

PIPER, Kelsey, «Death by Algorithm: The Age of Killer Robots Is Closer than You Think», *Vox*, 21 de junio de 2019. Disponible en: <https://www.vox.com/ 2019/6/21/18691459/killer-robots-lethal-autonomous-weapons-ai-war>.

POEL, Ibo van de, NIHLÉN, Jessica, FAHLQUIST, DOORN, Neelke, ZWART, Sjoerd D. y ROYAKKERS, Lambèr, «The Problem of Many Hands: Climate Change as an Example», *Science and Engineering Ethics*, 18, núm. 1 (2012), 49-67.

PRESS, Eyal, «The Wounds of the Drone Warrior», *New York Times Magazine*, 13 de junio de 2018. Disponible en: <https://www.nytimes.com/2018/ 06/13/magazine/veterans-ptsd-dro ne-warrior-wounds.html>.

PRICEWATERHOUSECOOPERS, *Fourth Industrial Revolution for the Earth: Harnessing Artificial Intelligence for the Earth*, PwC Network, 2018a.

—*Will Robots Really Steal Our Jobs? An International Analysis of the Potential Long Term Impact of Automation*, PwC UK y PwC Network, 2018b.

REESE, Hope, «Wi-Fi-Enabled “Hello Barbie” Records Conversations with Kids and Uses Al to Talk Back», *TechRepublic*, 10 de noviembre de 2015. Disponible en: <https://www.techrepublic.com/article/wi-fi-enabled-hello-barbie-records-conversations-with-kids-and-uses-ai-to-talk-back/>.

RICHARDSON, Kathleen, «The Asymmetrical “Relationship”: Parallels between Prostitution and the Development of Sex Robots», *ACM SIGCAS Computers and Society-Special Issue on Ethicomp*, 45, núm. 3 (2016), 290-293.

ROHRLICH, Justin, «Why a U.S. Soldier Turned against Drone Warfare», *Quartz*, 16 de octubre de 2019. Disponible en:< https://qz.com/1725819/ why-a-us-soldier-turned-against-drone-warfare/>.

RUDY-HILLER, Fernando, «The Epistemic Condition for Moral Responsibility», en Edward N. Zalta (ed.), *Stanford Encyclopedia of Philosophy*, otoño de 2018. Disponible en: <https://plato.stanford.edu/entries/moral-responsibility-epistemic/>.

Santoni de Sio, Filippo y Hoven, Jeroen van den, «Meaningful Human Control over Autonomous Systems: A Philosophical Account», *Frontiers in Robotics and Al* (28 de febrero de 2018). Disponible en: <https://www.frontiersin.org/ articles/10.3389/ frobt.2018.00015/fiill#B5>.

Schwab, Klaus, *The Fourth Industrial Revolution*, Nueva York, Crown Publishing Group, 2016.

Schomberg, Rene von (ed.), *Towards Responsible Research and Innovation in the Information and Communication Technologies and Security Technologies Fields*, Luxemburgo, Oficina de publicaciones de la Unión Europea, 2011. Disponible en: <http://ec.europa.eu/research/science-society/document-library/pdf_06/mep-rapport-2011_en.pdf>.

Sennett, Richard, *The Craftsman*, New Haven, CT, Yale University Press, 2008.

Servoz, Michel, *The Future of Work? Work of the Future! On How Artificial Intelligence, Robotics and Automation Are Transforming Jobs and the Economy in Europe*, Bruselas, Comisión Europea, 2019.

Sharkey, Amanda y Sharkey, Noel, «Granny and the Robots: Ethical Issues in Robot Care for the Elderly», *Ethics and Information Technology*, 14 (2012), 27-40.

Sharkey, Noel, «Killing Made Easy: From Joysticks to Politics», en Patrick Lin, George Bekey y Keith Abney (eds.), *Robot Ethics: The Ethical and Social Implications of Robotics*, Cambridge, MA, MIT Press, 2012, páginas 111-128.

Sharkey, Noel y Sharkey, Amanda, «The Crying Shame of Robot Nannies: An Ethical Appraisal», *Interaction Studies*, 11, núm. 2 (2010), 161-190.

Smith, Angela M., «Responsibility as Answerability», *Inquiry*, 58, núm. 2 (2015), 99-126.

Smith, Marquard, *The Erotic Doll: A Modern Fetish*, New Haven, CT, Yale University Press, 2013.

Sorell, Tom y Draper, Heather, «Robot Carers, Ethics, and Older People», *Ethics and Information Technology*, 16 (2014), 183-195.

Sparrow, Robert, «Robots in Aged Care: A Dystopian Future?», *Al and Society*, 31, núm. 4 (2016), 445-454.

—«Killer robots», *Journal of Applied Philosophy*, 24, núm. 1 (2007), 62-77.

Sparrow, Robert y Sparrow, Linda, «In the Hands of Machines? The Future of aged cared», *Minds and Machines*, 16, núm. 2 (2006), 141-161.

STAHL, Bernd C. y COECKELBERGH, Mark, «Ethics of Healthcare Robotics: Towards Responsible Research and Innovation», *Robotics and Autonomous Systems*, 86 (2016), 152-161.

SULLINS, John, «When Is a Robot a Moral Agent?», *International Review of Information Ethics*, 6 (2006), 23-30.

—«RoboWarfare: Can Robots Be More Ethical than Humans on the Battlefield?», *Ethics and Information Technology*, 12, núm. 3 (2010), 263-275.

—«Robots, Love, and Sex: The Ethics of Building a Love Machine», *IEEE Transactions on Affective Computing* 3, núm. 4 (2012), 398-409.

SUZUKI, Yutaka, GALLI, Lisa, IKEDA, Ayaka, ITAKURA, Shoji y KITAZAKI, Michiteru, «Measuring Empathy for Human and Robot Hand Pain Using Electroencephalography», *Scientific Reports*, 5 (2015), artículo núm. 15924. Disponible en: <https://doi.org/10.1038/srepl5924>.

TEGMARK, Max, *Life 3.0: Being Human in the Age of Artificial Intelligence*, Londres, Allen Lane, 2017.

TOGNAZZINI, Bruce, «Principles, Techniques, and Ethics of Stage Magic and Their Application to Human Interface Design», CHI ’93: Proceedings of the INTERACT ’93 y CHI ’93 *Conference on Human Factors in Computing Systems*, Nueva York, Association for Computing Machinery, 1993, págs. 355-362. Disponible en: <http://dl.acm.org/citation.cfm?id= 169284>.

TURKLE, Sherry, *Reclaiming Conversation: The Power of Talk in a Digital Age*, Nueva York, Penguin Books, 2015.

TURNER, Jacob, *Robot Rules: Regulating Artificial Intelligence*, Cham, Suiza, Palgrave Macmillan, 2019.

UNI GLOBAL UNION, *Top 10 Principles for Workers’ Data Privacy and Protection*, Nyon, Suiza, UNI Global Union. Disponible en: <http://www.thefutureworldofwork.org/media/35421/uni_workers_data_pro

VEAL, Anthony J., «The Leisure Society I: Myths and Misperceptions, 1960-79», *World Leisure Journal*, 53, núm. 3 (2011), 203-227.

VELIZ, Carissa, *«Privacy Is Power»: Why and How You Should Take Back Control of Your Data*, Londres, Penguin, 2020.

WAKEFIELD, Jane, «Foxconn Replaces “60,000 Factory Workers with Robots”», BBC, Technology, 25 de mayo de 2016. Disponible en: <https://www.bbc.com/news/technology-36376966>.

—«Can You Murder a Robot?», *BBC*, 17 de marzo de 2019. Disponible en: <https://www.bbc.com/news/technology-47090174>.

WALLACH, Wendell y ALLEN, Colin, *Moral Machines: Teaching Robots Right from Wrong*, Oxford, Oxford University Press, 2009.

WIENER, Norbert, *The Human Use of Human Beings: Cybernetics and Society*, Boston, Houghton Mifflin Co., 1954.

WILLIAMS, John, «Distant Intimacy: Space, Drones, and Just War», *Ethics and International Affairs*, 29, núm. 1 (2015), 93-110.

WINFIELD, Alan, «Making an Ethical Machine», *Open Transcripts* (blog). Acceso el 5 de enero de 2022. Disponible en: <http://opentranscripts.org/transcript/making-an-ethical-ma chine/>.

WINFIELD, Alan, BLUM, Christian y LIU, Wenguo, «Towards an Ethical Robot: Internal Models, Consequences and Ethical Action Selection», en Michael Mistry, Ales Leonardis, Mark Witkowski y Chris Melhuish (eds.), *Advances in Autonomous Robotics Systems. TAROS 2014. Lecture Notes in Computer Science*, Cham, Suiza, Springer, 2014, págs. 85-96.

WYNSBERGHE, Aimee van, «Designing Robots for Care: Care Centered ValueSensitive Design», *Science and Engineering Ethics*, 19, núm. 2 (2013), 407-433.

WYNSBERGHE, Aimee van y DONHAUSER, Justin, «The Dawning of the Ethics of Environmental Robots», *Science and Engineering Ethics*, 24, núm. 6 (2018), 1777-1800.

WYNSBERGHE, Aimee van y LI, Shuhong, «A Paradigm Shift for Robot Ethics: From HRI to Human-Robot-System Interaction (HRSI)», *Medicolegal and Bioethics*, 9 (2019), 11-21.

ZIMMERLI, Walther, «Deus Malignus», Artículo presentado en el Institut fur die Wissenschaften vom Menschen, Viena, octubre de 2019. Disponible en: <https://esel.at /termin/103418/walther-zimmerli-deus-malignus>.

ZUBOFF, Shoshana, *The Age of Surveillance Capitalism: The Fight for a Human Future at the New Frontier of Power*, Londres, Profile Books, 2019.

PARA SEGUIR LEYENDO

ASARO, Peter, «How Just Could a Robot War Be?», en Philip Brey, Adam Briggle y Katinka Waelbers (eds.), *Current Issues in Computing and Philosophy*, Amsterdam, IOS Press, 2008, págs. 50-64.

BENJAMIN, Ruha, *Race after Technology: Abolitionist Tools for the New Jim Code*, Cambridge, Reino Unido, Polity Press, 2019.

BOSTROM, Nick, *Superintelligence: Paths, Dangers, Strategies*, Oxford, Oxford University Press, 2014.

BRYNJOLFSSON, Erik y MCAFEE, Andrew, *The Second Machine Age: Work, Progress, and Prosperity in a Time of Brilliant Technologies*, Nueva York, W. W. Norton and Company, 2014.

BRYSON, Joanna, «Robots Should Be Slaves», en Yorick Wilks (ed.), *Close Engagements with Artificial Companions: Key Social, Psychological, Ethical and Design Issues*, Amsterdam, John Benjamins, 2010, págs. 63-74.

CARPENTER, Julie, «Deus Sex Machina: Loving Robot Sex Workers and the Allure of an Insincere Bliss», en John Danaher y Neil McArthur (eds.), *Robot Sex: Social and Ethical Implications*, Cambridge, MA, MIT Press, 2017, 261-287.

COECKELBERGH, *Growing Moral Relations: Critique of Moral Status Ascription*, Nueva York, Palgrave Macmillan, 2012.

—«The Moral Standing of Machines: Towards a Relational and NonCartesian Moral Hermeneutics», *Philosophy and Technology*, 27, núm. 1 (2014), 61-77.

—«AI, Responsibility Attribution, and a Relational Justification of Explainability», *Science and Engineering Ethics*, 26, núm. 4 (2020), 2051-2068.

—Mark, *Ética de la Inteligencia Artificial*, Madrid, Cátedra, 2021.

DANAHER, John, *Automation and Utopia: Human Flourishing in a World without Work*, Cambridge, MA, Harvard University Press, 2019.

FLORIDI, Luciano y SANDERS, J. W, «On the Morality of Artificial Agents», *Minds and Machines*, 14, núm. 3 (2004), 349-379.

FUNK, Michael, SEIBT, Johanna y COECKELBERGH, Mark, «Why Do/Should We Build Robots? Summary of a Plenary Discussion Session», en Mark Coeckelbergh, Janina Loh, Michael Funk, Johanna Seibt y Marco Norskov (eds.), *Envisioning Robots in Society: Power, Politics, and Public Space*, Actas de Robophilosophy 2018/TRANSOR 2018, University of Vienna, febrero de 2018, Amsterdam, IOS Press, 2018, págs. 369-384.

GUNKEL, David, «The Other Question: Can and Should Robots Have Rights?» *Ethics and Information Technology* 20, núm. 2 (2018), 87-99.

—*The Machine Question: Critical Perspectives on Al, Robots, and Ethics*. Cambridge, MA, MIT Press, 2012.

HARAWAY, Donna, «A Cyborg Manifesto», en *The Cybercultures Reader*, Londres, Routledge, 2000, en David Bell y Barbara M. Kennedy, págs. 291-324.

JOHNSON, Deborah, «Computer Systems: Moral Entities but Not Moral Agents», *Ethics and Information Technology* 8, núm. 4 (2006), 195-204.

LIN, Patrick, ABNEY, Keith y JENKINS, Ryan (eds.), *Robot Ethics 2.0: From Autonomous Cars to Artificial Intelligence*, Nueva York, Oxford University Press, 2017.

LIN, Patrick, BEKEY, George y ABNEY, Keith (eds.), *Robot Ethics: The Ethical and Social Implications of Robotics*, Cambridge, MA, MIT Press, 2012.

MATTHIAS, Andreas, «The Responsibility Gap: Ascribing Responsibility for the Actions of Learning Automata», *Ethics and Information Technology*, 6, núm. 3 (2004), 175-183.

NYHOLM, Sven, *Humans and Robots: Ethics, Agency, and Anthropocentrism*, Londres, Rowman and Littlefield, 2020.

SCHWAB, Klaus, *The Fourth Industrial Revolution*, Nueva York, Crown Publishing Group, 2016.

SHARKEY, Noel y SHARKEY, Amanda, «The Crying Shame of Robot Nannies: An Ethical Appraisal», *Interaction Studies*, 11, núm. 2 (2010), 161-190.

SPARROW, Robert y SPARROW, Linda, «In the Hands of Machines? The Future of Aged Care», *Minds and Machines*, 16, num. 2 (2006), 141-161.

SULLINS, John, «Robots, Love, and Sex: The Ethics of Building a Love Machine», *IEEE Transactions on Affective Computing*, 3, núm. 4 (2012), 398-409.

WYNSBERGHE, Aimee van, «Designing Robots for Care: Care Centered ValueSensitive Design», *Science and Engineering Ethics*, 19, núm. 2 (2013), 407-433.

WALLACH, Wendell y ALLEN, Colin, *Moral Machines: Teaching Robots Right from Wrong*, Oxford, Oxford University Press, 2009.

Índice analítico



MARK COECKELBERGH (Leuven, Bélgica, 1975). Filósofo de la tecnología, es profesor de filosofía de medios y tecnología en la Universidad de Viena.

Obras: *Liberation and Passion: Reconstructing the Passion Perspective on Human Being and Freedom*, 2002; *The Metaphysics of Autonomy: The Reconciliation of Ancient and Modern Ideals of the Person*, 2004; *Imagination and Principles: An Essay on the Role of Imagination in Moral Reasoning*, 2007; *Growing Moral Relations: Critique of Moral Status Ascription*, 2012; *Human Being @ Risk: Enhancement, Technology, and the Evaluation of Vulnerability Transformations*, 2013; *Environmental Skill: Motivation, Knowledge, and the Possibility of a Non-Romantic Environmental Ethics*, 2015; *Money Machines: Electronic Financial Technologies, Distancing, and Responsibility in Global Finance*, 2015; *New Romantic Cyborgs: Romanticism, Information Technology, and The End of the Machine*, 2017; *Moved by Machines: Performance Metaphors and Philosophy of Technology*, 2019; *Introduction to Philosophy of Technology*, 2019; *AI Ethics*, 2020; *Introduction to Philosophy of Technology*; Reijers, Wessel & Coeckelbergh, Mark, *Narrative and Technology Ethics*, 2020; *Self-improvement: technologies of the soul in the age of artificial intelligence*, 2022; *The political philosophy of AI: an introduction*, 2022; *Robot ethics*, 2022; *Digital Technologies, Temporality, and the Politics of Co-Existence*, 2023.

Notas

[1] Coeckelbergh, *New Romantic Cyborgs*. <<

[2] Disponible en: <https://robots.ieee.org/learn/>. <<

[3] Lin, Abney y Bekey, «Robots Ethics», 943. <<

[4] Heidegger, *Question concerning Technology*. <<

[5] Coeckelbergh, «You, Robot». <<

[6] Asaro, «What Should We Want from a Robot Ethic?», 9. <<

[7] Coeckelbergh, «Robotic Appearances». <<

[8] Abney, «Robotics, Ethical Theory, and Metaethics». <<

[9] Botting, *Mary Shelley*. <<

[10] Boddington, *Towards a Code of Ethics*; Coeckelbergh, *Ética de la Ingeligencia Artificial*; Gunkel, *An Introduction*; Liao, *Ethics*; Bartneck *et al*., *An Introduction*. <<

[11] Ver, por ejemplo, Hildebrandt, *Smart Technologies*; Turner, *Robot Rules*; Fosch-Villaronga, *Robots*. <<

[12] Lin, Bekey y Abney, *Robot Ethics*; Lin, Abney y Jenkins, *Robot Ethics 2.0*; Coeckelbergh *et al*., *Envisioning Robots*. <<

Capítulo 2

[1] Disponible en: <https://libcom.org/blog/xulizhi-foxconn-suicide-poetry>. <<

[2] Wakefield, «Foxconn». <<

[3] Marx, *Capital*, 548. <<

[4] Wiener, *Human Use*, 162. <<

[5] Servoz, *Future of Work*, 40. <<

[6] Fletcher y Webb, «Industrial Robot Ethics». <<

[7] UNI Global Union, *Top10Principles*, 4-5. <<

[8] Zuboff, *Age of Surveillance Capitalism*. <<

[9] Brynjolfsson y McAfee, *Second Machine Age*. <<

[10] Schwab, *Fourth Industrial Revolution*. <<

[11] Frey y Osborne, «Future of Employment». <<

[12] McKinsey Global Institute, *Jobs Lost*. <<

[13] *Ibídem*. <<

[14] Ford, *Rise of the Robots*. <<

[15] Frey y Osborne, «Future of Employment», 20-22. <<

[16] Servoz, *Future of Work*, 5. <<

[17] *Ibíd*., 39. <<

[18] *Ibíd*., 36. <<

[19] PricewaterhouseCoopers, *Will Robots Really Steal Our Jobs*? <<

[20] Servoz, *Future of Work*, 6. <<

[21] Foro Económico Mundial, *Reskilling Revolution*. <<

[22] PricewaterhouseCoopers, *Will Robots Really Steal Our Jobs?* <<

[23] *Ibidem*. <<

[24] McKinsey Global Institute, *Jobs Lost*. <<

[25] Servoz, *Future of Work*, 3. <<

[26] *Ibíd*., 45-46. <<

[27] Foro Económico Mundial, *Future of Jobs*, ix. <<

[28] Danaher, *Automation and Utopia*, 135. <<

[29] Veal, «Leisure». <<

[30] McKinsey Global Institute, *Jobs Lost*, 8. <<

[31] PricewaterhouseCoopers, *Fourth Industrial Revolution*. <<

Capítulo 3

[1] Disponible en: <https://www.jibo.com/>. <<

[2] Disponible en: <hnpsj7fúrurism.com/jibo-dead-announcement-video>. <<

[3] Ver la famosa demo de Googk Duplex en 2018: <https://ai.googleblog.com/2018/05/duplex-ai-system-for-naiiiatoüiirasation.-html>>. <<

[4] Para una concepción general de la discusión, ver Van Den Hoven *et al.*, «Privacy». <<

[5] Zuboff, *Age of Surveillance Capitalism*. <<

[6] Veliz, *Privacy Is Power*. <<

[7] Disponible en: <http://hellobarbiefaq.mattel.com/>. <<

[8] Reese, «WiFi-Enabled “Hello Barbie”». <<

[9] Calo, «Robots and Privacy», 188. <<

[10] Matsuzaki y Lindemann, «Autonomy-Safety-Paradox». <<

[11] Sparrow y Sparrow, «In the Hands of Machines?», 155. <<

[12] Sparrow, «Robots in Aged Care», 448, 446. <<

[13] Coeckelbergh, «Care Robots and the Future of ICT-Mediated Elderly Care». <<

[14] Sharkey y Sharkey, «Crying Shame», 166, 168. <<

[15] Turkle, *Reclaiming Conversation*, 344, 339, 358. <<

[16] Escribo «freudiana poco precisa» en lugar de «de Sigmund Freud», ya que mantengo distancias con su artículo de1927«Fetichismo» en el que perfila una teoría basada en sus opiniones específicas sobre el desarrollo sexual y ciertas patologías relacionadas. En el artículo, Freud considera el fetichismo una perversión masculina vinculada a los miedos del niño a la castración: confrontado con la ausencia de pene en la madre, el fetichismo busca un objeto como sustituto ante esa carencia. En los años 1950, Jacques-Marie-Emile Lacan desarrolló adicionalmente esta teoría. El significado de fetiche que uso aquí, por el contrario, es mucho más amplio y más común: se refiere a un objeto que es un sustituto de otra cosa. El término «fetiche» se usa, así, sin ninguna intención de evocar las teorías específicas de patología sexual desarrolladas por Freud y Lacan. Además, mi uso aquí del término «magia» se refiere al uso del término «fetiche» en el contexto de la antropología, pero, de nuevo, sin referirse a teorías específicas. Elaborar este tema nos distraería del objetivo introductorio de este libro. <<

[17] Sullins, «Robots, Love, and Sex», 208. <<

[18] Carpenter, «Deus Sex Machina», 283. <<

[19] Smith, *Erotic Doll*. <<

[20] Richardson, «Asymmetrical “Relationship”», 291, 292. Para la campaña, ver <https://campaignagainstsexrobots.org/>. <<

[21] Isaac y Bridewell, «White Lies», 157. <<

[22] Para más discusión sobre este tema, ver Coeckelbergh, «How to Describe». <<

[23] Coeckelbergh, «Technology Games/Gender Games». <<

[24] Carpenter *et al.*, «Gender Representation», 262. <<

[25] *Ibid*., 264. <<

[26] Benjamin, *Race after Technology*. <<

[27] Coeckelbergh, «Technology Games/Gender Games». <<

[28] Carpenter, «Existential Robot», 42. <<

Capítulo 4

[1] Foster, «Aging Japan». <<

[2] Disponible en: <http://www.parorobots.com/>. <<

[3] Archer, «“Friendly” Hospital Robot». <<

[4] Disponible en: <https://med.nyu.edu/robotic-surgery/physicians/what-robotic-surgery>. <<

[5] Mavroforou *et al.*, «Legal and Ethical Issues in Robotic Surgery». <<

[6] Nordrum, «Cyderdyne's HAL Exoskeleton». <<

[7] Van Wynsberghe y Li, «Paradigm Shift». <<

[8] Disponible en: <https://www.dream2020.eu/goals-methodology/>. <<

[9] Disponible en: <https://www.softbankrobotics.com/emea/en/nao>. <<

[10] Ver, por ejemplo, Coeckelbergh *et al.*, «Survey of Expectations». <<

[11] Sparrow y Sparrow, «In the Hands of Machines?»; Sharkey y Sharkey, «Granny». <<

[12] Matthias, «Responsibility Gap». <<

[13] Sparrow y Sparrow, «In the Hands of Machines?»; Sharkey y Sharkey, «Granny». <<

[14] Kitwood, *Dementia*, 47. Ver también Sharkey y Sharkey, «Granny». <<

[15] Sparrow y Sparrow, «In the Hands of Machines?». <<

[16] Coeckelbergh, «Health Care». <<

[17] Nozick, *Anarquía*, 43. <<

[18] Sharkey y Sharkey, «Granny». <<

[19] Beauchamp y Childress, *Principles*. <<

[20] Sparrow y Sparrow, «In the Hands of Machines?». <<

[21] Sharkey y Sharkey, «Granny». <<

[22] Nussbaum y Sen, *La calidad de la vida*; Nussbaum, *Las mujeres*; Coeckelbergh, «Health Care», 184-185; Coeckelbergh, «How I Learned». <<

[23] Nussbaum, *Las fronteras de la justicia, 76*. <<

[24] Nussbaum, *Ibid.*, *76-7$*. <<

[25] Coeckelbergh, «How I Learned», 79-80. <<

[26] Borenstein y Pearson, «Robot Caregivers». <<

[27] Coeckelbergh, «Care Robots, Virtual Virtue». <<

[28] Sparrow, «Robots in Aged Care», 448. <<

[29] Coeckelbergh, «E-Care as Craftmanship». <<

[30] Sennett, *Craftsman*. <<

[31] Dreyfus y Dreyfus, *Five-Stage Model*; Dreyfus y Dreyfus, «Towards a Phenomenology». <<

[32] Coeckelbergh, «E-Care as Craftmanship». <<

[33] Sorell y Draper, «Robot Carers», 188. <<

[34] Coeckelbergh, «Care Robots, Virtual Virtue». <<

[35] Coeckelbergh, «Artificial Agents», 267-270. <<

[36] Coeckelbergh *et al*., «Survey of Expectations». <<

[37] Sparrow y Sparrow, «In the Hands of Machines?». <<

[38] Friedman, «Value-Sensitive Design»; Friedman y Hendry, *Value Sensitive Design*. <<

[39] Van Wynsberghe, «Designing Robots». <<

[40] Von Schomberg, *Towards Responsible Research*. <<

[41] Stahl y Coeckelbergh, «Ethics». <<

Capítulo 5

[1] Disponible en: <http://moralmachine.mit.edu/>. <<

[2] Awad *et al.*, «Moral Machine». <<

[3] Nyholm y Smids, «Ethics of Accident-Algorithms». <<

[4] Levin y Wong, «Self-Driving Uber». <<

[5] Anderson y Anderson, *Machine Ethics*, 9. <<

[6] Winfield, «Making an Ethical Machine». <<

[7] Asimov, «Runaround». <<

[8] Winfield, Blum y Liu, «Towards an Ethical Robot». <<

[9] Wallach y Allen, *Moral Machines*, 7. <<

[10] Johnson, «Computer Systems». <<

[11] Coeckelbergh, «Moral Appearances». <<

[12] Clark, «Asimovs Laws». <<

[13] Bostrom, *Superintelligence*, 139. <<

[14] Wallach y Allen, *Moral Machines*, 8, 9, 26. <<

[15] Floridi y Sanders, «On the Morality». <<

[16] Sullins, «When Is a Robot a Moral Agent?», 29. <<

[17] Nyholm, *Humans and Robots*, 55. <<

[18] Helveke y Nida-Rümelin, «Responsibility for Crashes». <<

[19] Matthias, «Responsibility Gap». <<

[20] Fischer y Ravizza, *Responsibility and Control*; Rudy-Hiller, «Epistemic Condition». <<

[21] Aristóteles, *Obras completas*, 1109b30-1111b5. <<

[22] Van de Poel *et al.*, «Many Hands». <<

[23] Fischer y Ravizza, *Responsibility and Control*, 13. <<

[24] Aristóteles, *Obras completas*, Illla3-lllla5. <<

[25] Rudy-Hiller, «Epistemic Condition». <<

[26] Duff, «Who Is Responsible?». <<

[27] Smith, «Responsibility as Answerability». <<

[28] Coeckelbergh, «AI, Responsability Attribution». <<

[29] Coeckelbergh, «AI, Responsability Attribution». <<

[30] Loh y Loh, «Autonomy and Responsability», 36. <<

[31] Coeckelbergh, «Virtual Moral Agency». <<

[32] Sullins, «When Is a Robot a Moral Agent?». <<

Capítulo 6

[1] Parke, «Is it Cruel?» <<

[2] Leopold, «HitchBOT». <<

[3] Wakefield, «Can You Murder». <<

[4] Suzuki *et al.*, «Measuring Empathy». <<

[5] Darling, «Who’s Johnny», 181. <<

[6] Horstmann *et al.*, «Do a Robots Social Skills and Its Objection Discourage Interactants?», 20-21. <<

[7] Ishiguro, «Android Science». <<

[8] Mori, «Uncanny Valley»; MacDorman e Ishiguro, «Uncanny Advantage».
<<

[9] Tognazzini, «Principies, Techniques, and Ethics of Stage Magic». <<

[10] Flusser, *Shape of Things*, 17. <<

[11] Coeckelbergh, «How to Describe». <<

[12] Bryson, «Robots Should Be Slaves». <<

[13] Gunkel, «Other Question». <<

[14] Kant, *Lecciones de ética*. <<

[15] Darling, «Extending Legal Protection». <<

[16] Coeckelbergh, «You, Robot». <<

[17] Coeckelbergh, *Growing Moral Relations*; Coeckelbergh, «Moral Standing of Machines». <<

[18] Coeckelbergh, *Growing Moral Relations* <<

[19] Gunkel, *Machine Question*; Gunkel, «Other Question». <<

[20] Ver, por ejemplo, Danaher, «Welcoming Robots»; Coeckelbergh, «Should We Treat Teddy Bear 2.0 as a Kantian Dog?» . <<

[21] Coeckelbergh, «Why Care about Robots?». <<

[22] Coeckelbergh, *Moved by Machines*. <<

[23] Coeckelbergh, «How to Describe»; Zimmerli, «Deus Malignus». <<

[24] Danaher, *Automation and Utopia*, 186. <<

[25] Carpenter, «Deus Sex Machina». <<

[26] Coeckelbergh, «How to Use»; Coeckelbergh, «Should We Treat Teddy Bear 2.0 as a Kantian Dog?». <<

Capítulo 7

[1] Nótese que el término «no tripulado» es problemático en términos de género (*N. del T.* en inglés, *unmanned*, literalmente, que no está controlado por un hombre). <<

[2] Piper, «Death by Algorithm». <<

[3] Disponible en: <https://www.stopkillerrobots.org/>. <<

[4] McKay, «19 Climate Change Activists». <<

[5] Lin, Bekey y Abney, *Autonomous Military Robotics*. <<

[6] Kreps, «Ground the Drones?». <<

[7] Arkin, «Case for Ethical Autonomy». <<

[8] Asaro, «What Should We Want from a Robot Ethic?». <<

[9] Müller, «Autonomous Killer Robots», 76. <<

[10] Sullins, «RoboWarfare». <<

[11] Byrne, «Making Drones». <<

[12] Sparrow, «Killer Robots», 70, 71, 75. <<

[13] Müller, «Autonomous Killer Robots», 75-76, 78. <<

[14] Ver, por ejemplo, *Article 36*, «Killing by Machine». <<

[15] Santoni de Sio y Van den Hoven, «Meaningful Human Control». <<

[16] Ver, por ejemplo, Gross, «Assassination». <<

[17] Heyns, *Report*, 20-21. <<

[18] Grossman, *On Killing*. <<

[19] Sharkey, «Killing Made Easy». <<

[20] Coeckelbergh, «Drones, Information Technology, and Distance». <<

[21] Linebaugh, «I Worked on the U. S. Drone Program». <<

[22] Bumiller, «Dayjob»; Coeckelbergh, «Drones, Information Technology, and Distance»; Coeckelbergh, «Drones, Morality, and Vulnerability». <<

[23] Williams, «Distant Intimacy»; Coeckelbergh, «Drones, Information Technology, and Distance». <<

[24] McCammon, «Warfare». <<

[25] Press, «Wounds». <<

[26] Monbiot, «With Its Deadly Drones». <<

[27] Enemark, *Armed Drones*, 7, 4. <<

[28] Coeckelbergh, «Drones, Morality, and Vulnerability». <<

[29] Heyns, *Report*, 17. <<

[30] Maller, «Secretary-General’s Message». <<

[31] Campaña para detener a los robots asesinos, «Parliamentary Actions in Europe». <<

Capítulo 8

[1] Disponible en: <https://conferences.au.dk/robo-philosophy-2020-at-aarhus-university/ previous-conferences/rp2018 / >. <<

[2] Funk, Seibt y Coeckelbergh, «Why Do/Should We Build Robots?». <<

[3] Descartes, *Discurso del método*, parte 5. <<

[4] Coeckelbergh, *Human Being @ Risk*; Coeckelbergh, «Robotic Appearances». <<

[5] Moravec, *Mind Children*, 1. <<

[6] Kurzweil, *Singularity Is Near*. <<

[7] Bostrom, *Superintelligence*. <<

[8] Tegmark, *Life 3.0*, 335. <<

[9] Haraway, «Cyborg Manifestó», 292, 295. <<

[10] Latour, *Nunca fuimos modernos*. <<

[11] Braidotti, *Posthuman*, 3, 43, 90, 92, 94, 102. <<

[12] Van Wynsberghe y Donhauser, «Dawning». <<

Bibliografía

[*] Las fechas de última consulta de los hipervínculos que figuran a lo largo de este libro se corresponden con los meses de preparación de la edición original, publicada por The MIT Press en septiembre de 2022. <<

LA ÉTICA DE LOS ROBOTS

Mark Coeckelbergh

